





Article

The Multi-Domain International Search on Speech 2020 ALBAYZIN Evaluation: Overview, Systems, Results, Discussion and Post-Evaluation Analyses

Javier Tejedor ^{1,*} , Doroteo T. Toledano ² , Jose M. Ramirez ³, Ana R. Montalvo ³ 
and Juan Ignacio Alvarez-Trejos ² 

- ¹ Institute of Technology, Universidad San Pablo-CEU, CEU Universities, Urbanización Montepríncipe, 28668 Boadilla del Monte, Spain
- ² AUDIAS, Electronic and Communication Technology Department, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Av. Francisco Tomás y Valiente, 11, 28049 Madrid, Spain; doroteo.torre@uam.es (D.T.T.); juani.alvarez@estudiante.uam.es (J.I.A.-T.)
- ³ Voice Group, Advanced Technologies Application Center, CENATAV, Rpto. Siboney, Playa, La Habana 74390, Cuba; jsanchez@cenatav.co.cu (J.M.R.); amontalvo@cenatav.co.cu (A.R.M.)
- * Correspondence: javier.tejedor@ceu.es

Abstract: The large amount of information stored in audio and video repositories makes search on speech (SoS) a challenging area that is continuously receiving much interest. Within SoS, spoken term detection (STD) aims to retrieve speech data given a text-based representation of a search query (which can include one or more words). On the other hand, query-by-example spoken term detection (QbE STD) aims to retrieve speech data given an acoustic representation of a search query. This is the first paper that presents an internationally open multi-domain evaluation for SoS in Spanish that includes both STD and QbE STD tasks. The evaluation was carefully designed so that several post-evaluation analyses of the main results could be carried out. The evaluation tasks aim to retrieve the speech files that contain the queries, providing their start and end times and a score that reflects how likely the detection within the given time intervals and speech file is. Three different speech databases in Spanish that comprise different domains were employed in the evaluation: the MAVIR database, which comprises a set of talks from workshops; the RTVE database, which includes broadcast news programs; and the SPARL20 database, which contains Spanish parliament sessions. We present the evaluation itself, the three databases, the evaluation metric, the systems submitted to the evaluation, the evaluation results and some detailed post-evaluation analyses based on specific query properties (in-vocabulary/out-of-vocabulary queries, single-word/multi-word queries and native/foreign queries). The most novel features of the submitted systems are a data augmentation technique for the STD task and an end-to-end system for the QbE STD task. The obtained results suggest that there is clearly room for improvement in the SoS task and that performance is highly sensitive to changes in the data domain.

Keywords: search on speech; spoken term detection; query-by-example spoken term detection; international evaluation; Spanish language



Citation: Tejedor, J.; Toledano, D.T.; Ramirez, J.M.; Montalvo, A.R.; Alvarez-Trejos, J.I. The Multi-Domain International Search on Speech 2020 ALBAYZIN Evaluation: Overview, Systems, Results, Discussion and Post-Evaluation Analyses. *Appl. Sci.* **2021**, *11*, 8519. <https://doi.org/10.3390/app11188519>

Academic Editor: Valentín Cardeñoso-Payo

Received: 30 July 2021

Accepted: 10 September 2021

Published: 14 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The huge amount of information stored in audio and audiovisual repositories makes it necessary to develop efficient methods for search on speech (SoS). Significant research has been carried out for years in this area, and, in particular, in the tasks of spoken document retrieval (SDR) [1–6], keyword spotting (KWS) [7–13], spoken term detection (STD) [14–25] and query-by-example spoken term detection (QbE STD) [26–31].

1.1. Spoken Term Detection Overview

Spoken term detection aims to find *terms* within audio archives. This is based on a text-based input, commonly the word/phone transcription of the search term, and hence STD is also called text-based STD.

Spoken term detection systems typically comprise three different stages: (1) the audio is decoded into word/subword lattices using an automatic speech recognition (ASR) subsystem trained for the target language; (2) a term detection subsystem searches the terms within those word/subword lattices to hypothesise detections; and (3) confidence measures are computed to rank the detections. The STD systems are normally language-dependent and require large amounts of language resources.

1.2. Query-by-Example Spoken Term Detection Overview

Query-by-example spoken term detection also aims to search within audio archives, but it is based on an acoustic (spoken) input. In QbE STD, we consider the scenario in which the user finds a segment of speech that contains terms of interest within a speech data repository, and the user's purpose is to find similar speech segments within that repository. The speech segment found is the query and the system outputs other similar segments from the repository, which we will henceforth refer to as utterances. Alternatively, the query can be uttered by the user. This is a highly valuable task for blind people or devices that do not have a text-based input, and, consequently, the query must be given in another format, such as speech.

Query-by-example spoken term detection has been traditionally addressed using three different approaches: methods based on the word/subword transcription of the query, methods based on template matching of features and hybrid approaches. In the last couple of years, however, there have been new proposals based on deep learning.

1.3. Difference between Spoken Term Detection and Query-by-Example Spoken Term Detection

It should be noted that STD and QbE STD tasks are quite similar. Both tasks aim to retrieve the speech files that contain the query of interest along with the appropriate timestamps. They only differ in the input format. Whereas, for STD, the input is text, for QbE STD it is speech. Although it is true that an STD system can be used to build a QbE STD system by decoding the acoustic query with an ASR system and then performing an text-based STD search, both tasks are fully-independent.

1.4. Related Work

In this section, we summarise the previous work related to both STD and QbE STD tasks.

1.4.1. Spoken Term Detection

Spoken term detection has been a hot topic in the past few years due to its many applications and has received a great deal of interest from many outstanding companies and research institutes, such as IBM [14,32–36], BBN [37–39], SRI & OGI [40–42], BUT [17,43,44], Microsoft [45], QUT [46,47], JHU [16,48–50], Fraunhofer IAIS/NTNU/TUD [15], NTU [31,51], IDIAP [52] and Google [21], among others. Within an STD system, the ASR subsystem uses mostly word-based speech recognition [24,41,53–59] due to its better performance in comparison with subword-based approaches.

However, subword-based ASR [22,24,60–66] is also being used, sometimes in combination with word-based ASR. One of the main challenges of using word-based ASR in this context is that, in principle, only in-vocabulary (INV) terms can be detected. Subword-based ASR, on the other hand, can detect terms even though they are not in the vocabulary of the recognizer (i.e., out-of-vocabulary (OOV) terms). A more robust system could be obtained by combining both approaches [17,24,25,38,39,42,50,67–73].

The availability of ASR toolkits, such as Kaldi [50,74] and ESPnet [75], among others, facilitates the development of STD systems. For instance, Kaldi includes the tools to build a complete STD system since it integrates an ASR subsystem, a term detector and a decision

maker [74,76,77]. It uses a word-based approach and proposes a method based on proxy-words for OOV detection that replaces each OOV word by the most similar in-vocabulary word (or word sequence) [78].

In current state-of-the-art neural end-to-end ASR approaches, it is common to use plain characters or word fragments as units [75,79], which theoretically avoids the OOV issue. However, in practice, these systems are combined with language models to improve recognition accuracy, which again limits the vocabulary of the ASR system.

Deep learning and, in particular, end-to-end systems were also recently investigated to solve the STD problem directly. In this direction, several end-to-end ASR-free approaches for STD were proposed [13,34–36]. In addition to exploring neural end-to-end approaches, deep learning is extensively used to extract representations (embeddings) of audio documents and query terms that facilitate the search [20,21,23,25].

1.4.2. Query-by-Example Spoken Term Detection

It is possible to use a text-based STD system to solve the problem of QbE STD by first transcribing the acoustic query, which can be done automatically using an ASR system and, thus, converting the QbE STD problem into a text-based STD problem. However, errors produced in the transcription of the query are difficult to recover and lead to a significant performance degradation. Some examples of this approach using different models and units are [26–28,80–88]. More recently, this approach has been extended with automatic unit discovery [89–91] and deep neural networks (DNNs) for extracting bottleneck features [92,93].

Another approach, perhaps the most common one, transforms audio (both queries and utterances) into sequences of acoustic features and then makes use of template matching methods to find subsequences in the utterance representations that are similar to the sequence representing the query. This approach typically outperforms transcription-based techniques in QbE STD [94]. In addition, this approach can lead to language-independent STD systems, since prior knowledge of the language is not needed. The two main variations of this approach are the features used to represent the audio and the template matching method used.

The most commonly used features are posteriorgrams, and, among them, the most common type is the phoneme posteriorgram [87,95–104]. Another frequent type is the Gaussian posteriorgram [29,82,95,105,106]. Finally, there are also works exploring other types of posteriorgrams [107–110]. In recent years, the use of bottleneck features extracted from DNNs became popular [87,92,93,111–115].

There are also methods that explore other types of features beyond posteriorgrams and bottlenecks [116–119]. Most of the previous works make use of dynamic time warping (DTW) for query search, in many cases in the form of subsequence DTW (S-DTW) and some variants [29,95,100–103,105,108,110,112,115,120]. A well-known problem of these methods is the computational cost. To reduce this cost, Ref. [104] proposes hashing the phone posteriors to speed up the search and thus enable searching on massively large datasets.

Aiming to keep the advantages of both methods, hybrid methods that combine the two previous approaches were proposed. Most works use logistic regression-based fusion of DTW and phoneme or syllable-based speech recognition systems [121–127]. Other hybrid approaches combine DTW with other techniques, such as subspace modelling [110]. Some more recent approaches use DNNs for posteriorgram-based rescoring [128,129].

In the last few years, there were a few novel proposals in QbE STD, most of them based on deep learning. The most direct use of deep learning is perhaps the one proposed in [130,131], where a convolutional neural network (CNN) is used to decide whether the DTW cost matrix (considered as a grey scale image) contains a match. A less direct approach consists of replicating the standard approach of natural language processing (NLP) of representing a word with a fixed-length vector (embedding). In [120,132–134], this is extended by obtaining the word embedding directly from the audio. Once the embeddings are obtained, matching words is trivial and can be done using nearest neighbours [132].

Finally, attention mechanisms in DNNs allow the system to focus on the parts of the audio that are more relevant.

This was applied in the context of QbE STD to train a neural end-to-end system in which the attention weights indicate the time span of the detected queries [135]. More recently, attention mechanisms were extended to two-way attention mechanisms in the NLP context of question answering [136], and this approach was also applied to the problem of QbE STD [137].

2. Search on Speech Evaluation

2.1. Evaluation Summary

The SoS evaluation involves searching a list of queries (given in both written and acoustic forms) within speech data. As a result, both audio files and timestamps of each detected occurrence must be given.

Specifically, the evaluation consists in searching different queries within different sets of speech data. Workshop talks, broadcast news and parliament sessions domains are considered in the evaluation. Individual speech and text datasets are provided for each domain. Each domain contains training/development/test data, except for the parliament sessions dataset, for which only test data are provided. The evaluation results and rankings are established independently for each domain using the corresponding test data.

Regarding the system construction, participants are allowed to use the training data only for system training and the development data only for system tuning; however, any additional data can also be employed both for system training and development.

Two different types of queries are defined in this evaluation: in-vocabulary queries and out-of-vocabulary queries. The OOV query set aims to simulate the out-of-vocabulary words of a large vocabulary continuous speech recognition (LVCSR) system. In the case where participants employ an LVCSR system for system construction, these OOV words should be previously removed from the system dictionary, and hence other methods (e.g., phone-based systems) need to be used for searching OOV queries. The only exception is for end-to-end system construction, for which participants are allowed to treat all the queries as INV queries. On the other hand, the words present in the INV queries could appear in the LVCSR system dictionary if the participants find it suitable.

For the QbE STD task, participants are allowed to make use of the target language information (Spanish) when building their system/s (i.e., system/s can be language-dependent), although evaluation organizers highly encouraged participants to develop language-independent systems.

Participants can submit a primary system and up to four contrastive systems for each task. No manual intervention is allowed for each system to generate the final output file, and hence all systems have to be fully automatic. Listening to the test data or any other human interaction with the test data is forbidden before the evaluation results are sent back to the participants. The standard extensible markup language (XML)-based format accepted by the national institute of standards and technology (NIST) evaluation tool [138] is used for query detection. Test data ground-truth labels are given to participants once organizers send them back the evaluation results.

About five months were given to participants for system development, and therefore the SoS evaluation focuses on building SoS systems in a limited period of time. Training and development data were released by mid-March 2020. Test data were released by the beginning of September 2020. System submission was due by mid-October 2020. The final results were discussed at IberSPEECH 2020 conference by the end of March 2021.

2.2. Databases

Three databases that comprise different acoustic conditions and domains were employed for the evaluation: the workshop talks MAVIR and broadcast news RTVE databases, which were used in previous ALBAYZIN SoS evaluations, and the SPARL20 database, which was the new one added for this evaluation and which contains speech from Spanish

parliament sessions held from 2016. For the MAVIR and RTVE databases, three independent datasets (i.e., training, development and testing) were provided to participants.

For the SPARL20 database, only test data were provided. This allowed measuring the system generalization capability to an unseen domain. Tables 1–3 show some of the database features, such as the training/development/test division, the number of word occurrences, the duration, the number of speakers and the average mean opinion score (MOS) [139]. The latter is included to show the quality of each speech file in the databases.

Table 1. Characteristics of the MAVIR database: number of word occurrences (#occ.), duration (dur.) in minutes (min), number of speakers (#spk.) and average MOS (Ave. MOS). These characteristics are displayed for training (train), development (dev) and testing (test) datasets.

| File ID | Data | #occ. | dur. (min) | #spk. | Ave. MOS |
|----------|-------|--------|------------|-------------------|----------|
| Mavir-02 | train | 13,432 | 74.51 | 7 (7 ma.) | 2.69 |
| Mavir-03 | dev | 6681 | 38.18 | 2 (1 ma. 1 fe.) | 2.83 |
| Mavir-06 | train | 4332 | 29.15 | 3 (2 ma. 1 fe.) | 2.89 |
| Mavir-07 | dev | 3831 | 21.78 | 2 (2 ma.) | 3.26 |
| Mavir-08 | train | 3356 | 18.90 | 1 (1 ma.) | 3.13 |
| Mavir-09 | train | 11,179 | 70.05 | 1 (1 ma.) | 2.39 |
| Mavir-12 | train | 11,168 | 67.66 | 1 (1 ma.) | 2.32 |
| Mavir-04 | test | 9310 | 57.36 | 4 (3 ma. 1 fe.) | 2.85 |
| Mavir-11 | test | 3130 | 20.33 | 1 (1 ma.) | 2.46 |
| Mavir-13 | test | 7837 | 43.61 | 1 (1 ma.) | 2.48 |
| ALL | train | 43,467 | 260.27 | 13 (12 ma. 1 fe.) | 2.56 |
| ALL | dev | 10,512 | 59.96 | 4 (3 ma. 1 fe.) | 2.64 |
| ALL | test | 20,277 | 121.3 | 6 (5 ma. 1 fe.) | 2.65 |

Table 2. Characteristics of the RTVE database: number of word occurrences (#occ.), duration (dur.) in minutes (min), number of speakers (#spk.) and average MOS (Ave. MOS). These characteristics are displayed for training (train), development (dev), and testing (test) datasets. The results for train and dev1 are not reported per file due to the large number of files (about 400 for train and about 60 for dev1).

| File ID | Data | #occ. | dur. (min) | #spk. | Ave. MOS |
|---------------------|------|--------|------------|-------|----------|
| LN24H-20151125 | dev2 | 21,049 | 123.50 | 22 | 3.37 |
| LN24H-20151201 | dev2 | 19,727 | 112.43 | 16 | 3.27 |
| LN24H-20160112 | dev2 | 18,617 | 110.40 | 19 | 3.24 |
| LN24H-20160121 | dev2 | 18,215 | 120.33 | 18 | 2.93 |
| millennium-20170522 | dev2 | 8330 | 56.50 | 9 | 3.61 |
| millennium-20170529 | dev2 | 8812 | 57.95 | 10 | 3.24 |
| millennium-20170626 | dev2 | 7976 | 55.68 | 14 | 3.55 |
| millennium-20171009 | dev2 | 9863 | 58.78 | 12 | 3.60 |
| millennium-20171106 | dev2 | 8498 | 59.57 | 16 | 3.40 |
| millennium-20171204 | dev2 | 9280 | 60.25 | 10 | 3.29 |
| millennium-20171211 | dev2 | 9502 | 59.70 | 12 | 2.95 |
| millennium-20171218 | dev2 | 9386 | 55.55 | 15 | 2.70 |
| EC-20170513 | test | 3565 | 22.13 | N/A | 3.12 |

Table 2. *Cont.*

| File ID | Data | #occ. | dur. (min) | #spk. | Ave. MOS |
|---------------------|-------|-----------|------------|-------|----------|
| EC-20170520 | test | 3266 | 21.25 | N/A | 3.38 |
| EC-20170527 | test | 2602 | 17.87 | N/A | 3.42 |
| EC-20170603 | test | 3527 | 23.87 | N/A | 3.90 |
| EC-20170610 | test | 3846 | 24.22 | N/A | 3.31 |
| EC-20170617 | test | 3368 | 21.55 | N/A | 3.36 |
| EC-20170624 | test | 3286 | 22.60 | N/A | 3.65 |
| EC-20170701 | test | 2893 | 22.52 | N/A | 3.47 |
| EC-20170708 | test | 3425 | 23.15 | N/A | 3.58 |
| EC-20170715 | test | 3316 | 22.55 | N/A | 3.82 |
| EC-20170722 | test | 3929 | 27.40 | N/A | 3.88 |
| EC-20170729 | test | 4126 | 27.45 | N/A | 3.61 |
| EC-20170909 | test | 3063 | 21.05 | N/A | 3.64 |
| EC-20170916 | test | 3422 | 24.60 | N/A | 3.40 |
| EC-20170923 | test | 3331 | 22.02 | N/A | 3.24 |
| EC-20180113 | test | 2742 | 19.02 | N/A | 3.80 |
| EC-20180120 | test | 3466 | 21.97 | N/A | 3.28 |
| EC-20180127 | test | 3488 | 22.52 | N/A | 3.56 |
| EC-20180203 | test | 3016 | 21.60 | N/A | 3.90 |
| EC-20180210 | test | 3214 | 23.20 | N/A | 3.71 |
| EC-20180217 | test | 3094 | 20.33 | N/A | 3.57 |
| EC-20180224 | test | 3140 | 20.78 | N/A | 3.56 |
| millennium-20170703 | test | 8714 | 55.78 | N/A | 1.10 |
| millennium-20171030 | test | 8182 | 57.05 | N/A | 3.44 |
| ALL | train | 3,729,924 | 27729 | N/A | 3.04 |
| ALL | dev1 | 545,952 | 3742.88 | N/A | 2.90 |
| ALL | dev2 | 149,255 | 930.64 | N/A | 3.25 |
| ALL | test | 90,021 | 605.48 | N/A | 3.32 |

Table 3. Characteristics of the SPARL20 database, used as test data in the evaluation: number of word occurrences (#occ.), duration (dur.) in minutes (min), number of speakers (#spk.) and average MOS (Ave. MOS). These characteristics are displayed for training (train), development (dev) and testing (test) datasets.

| File ID | #occ. | dur. (min) | #spk. | Ave. MOS |
|------------------------------|-------|------------|-----------------|----------|
| 13_000500_003_1_19421_642906 | 875 | 5.55 | 2 (1 ma. 1 fe.) | 3.11 |
| 13_000400_007_0_19432_643097 | 563 | 3.53 | 2 (1 ma. 1 fe.) | 3.47 |
| 13_000400_005_0_19422_642932 | 718 | 3.57 | 2 (1 ma. 1 fe.) | 2.92 |
| 13_000400_005_0_19422_642923 | 1898 | 11.62 | 1 (1 fe.) | 3.27 |
| 13_000400_005_0_19422_642922 | 1733 | 11.67 | 1 (1 fe.) | 3.19 |
| 13_000400_004_0_19388_642448 | 1107 | 7.43 | 1 (1 ma.) | 2.53 |
| 13_000400_003_0_19381_642399 | 1403 | 8.13 | 3 (2 ma. 1 fe.) | 2.83 |

Table 3. Cont.

| File ID | #occ. | dur. (min) | #spk. | Ave. MOS |
|------------------------------|--------|------------|-------------------|----------|
| 13_000400_003_0_19381_642398 | 1279 | 11.45 | 3 (2 ma. 1 fe.) | 3.26 |
| 13_000400_002_1_19376_642375 | 2007 | 13.70 | 2 (1 ma. 1 fe.) | 2.41 |
| 13_000400_002_1_19376_642366 | 1720 | 10.73 | 1 (1 ma.) | 2.27 |
| 13_000327_002_0_19437_643241 | 1405 | 8.73 | 2 (2 ma.) | 3.37 |
| 12_000400_153_0_18748_633006 | 1331 | 8.33 | 2 (2 ma.) | 3.48 |
| 12_000400_148_0_18727_632388 | 1012 | 5.42 | 2 (1 ma. 1 fe.) | 3.14 |
| 12_000400_003_0_16430_586456 | 1484 | 10.33 | 1 (1 ma.) | 2.17 |
| ALL | 18,535 | 120.19 | 25 (16 ma. 9 fe.) | 2.90 |

2.2.1. MAVIR

The MAVIR database consists of a set of Spanish talks from the MAVIR workshops (<http://www.mavir.net>, accessed on 10 september 2021) held in 2006, 2007 and 2008. It contains speech from Spanish speakers both from Spain and Latin America.

The MAVIR Spanish data consist of 7 h of spontaneous speech files from different speakers. These data were then divided into training, development and test sets. The data were manually annotated in an orthographic form, but timestamps were only set for phrase boundaries. For the SoS evaluation, organizers manually added the timestamps for the roughly 3000 occurrences of the queries used in the development and test evaluation sets. The training data were made available to the participants, including the orthographic transcription and the timestamps for phrase boundaries (<http://cartago.llf.uam.es/mavir/index.pl?m=videos>, accessed on 10 september 2021).

Initially, the speech data were recorded in several audio formats (pulse code modulation (PCM) mono and stereo, MP3, 22.05 kHz, 48 kHz, among others). Recordings were afterward converted to PCM, 16 kHz, single channel, 16 bits per sample using the SoX tool (<http://sox.sourceforge.net/>, accessed on 10 september 2021) for this evaluation. All the recordings except for one were originally made with a Digital TASCAM DAT model DA-P1 equipment. Different microphones were used, which mainly consisted of tabletop or floor standing microphones, and one lavalier microphone was also employed. The distance from the microphone to the mouth of the speaker was not specifically controlled, but in most of the cases was smaller than 50 cm.

The speech recordings took place in large conference rooms with capacity for over a hundred people. This conveys additional challenges including background noise (particularly babble noise) and reverberation. Therefore, these realistic settings and the variety of phenomena in the spontaneous speech make this database appealing and challenging enough for the SoS evaluation.

2.2.2. RTVE

The RTVE database belongs to the broadcast news domain and contains speech from different television (TV) programs recorded from 2015 to 2018 (e.g., Millenium, Al filo de lo imposible, Asuntos públicos, La tarde en 24H, to name a few). These amount to about 570 h in total, which were further divided into training, development and test sets. To prepare the data for the evaluation, organizers manually added the timestamps for the roughly 2700 occurrences of the queries used in the development and test evaluation sets.

The training speech data along with the corresponding subtitles (even though these could contain non-accurate word transcriptions) were provided to participants. The development data were divided into two different development sets, as follows: The *dev1* dataset consists of about 60 h of speech and human-revised word transcriptions without time alignment. The *dev2* dataset, which was employed as *real* development data for the SoS evaluation, consists of 15 h of speech data. The format of the recordings is advanced

audio coding (AAC), stereo, 44.1 kHz and variable bit rate. More information about the RTVE database can be found in [140].

2.2.3. SPARL20

The SPARL20 database consists of a small subset of speech from Spanish parliament sessions held from 2016. The SPARL20 data consist of spontaneous speech and amount to about 2 h of speech extracted from 14 audio files. For the SoS evaluation, organizers manually added the timestamps for the roughly 1500 occurrences of the queries used as test data.

The original recordings are videos in moving picture experts group (MPEG) format. The evaluation organizers extracted the audio of these videos and converted them to PCM, 16 kHz, single channel and 16 bits per sample using the *ffmpeg* tool (<https://ffmpeg.org/>, accessed on 10 september 2021). This database contains several noise types (e.g., laugh, applause, etc.), which makes it quite challenging.

2.2.4. Query List Selection

Query selection plays an important role within search on speech, since it should carefully take into account different search scenarios. To do so, the queries involved both in the STD and the QbE STD tasks include high occurrence queries, low occurrence queries, in-language (INL) (i.e., Spanish) queries, out-of-language (OOL) (i.e., foreign) queries, single-word and multi-word queries, in-vocabulary and out-of-vocabulary queries and queries of different length. In the evaluation datasets, a query may not have any occurrence or appear one or more times in the speech data. Table 4 includes some features of the development and test query lists, such as the number of INL and OOL queries, the number of single-word and multi-word queries and the number of INV and OOV queries, along with the number of occurrences of each set in the corresponding speech database for the STD task.

Table 4. Development and test query list characteristics for the MAVIR, RTVE and SPARL20 databases for the STD task. ‘dev’ stands for development, ‘INL’ refers to in-language queries, ‘OOL’ to foreign terms, ‘SING’ to single-word queries, ‘MULTI’ to multi-word queries, ‘INV’ to in-vocabulary queries, ‘OOV’ to out-of-vocabulary queries and ‘occ.’ stands for occurrences. The term length of the development query lists varies between 4 and 27 graphemes. The term length of the MAVIR and RTVE test query lists varies between 4 and 28 graphemes. The term length of the SPARL20 test query list varies between 3 and 19 graphemes.

| Query List | Dev-MAVIR | Dev-RTVE | Test-MAVIR | Test-RTVE | Test-SPARL20 |
|---------------|-----------|------------|------------|------------|--------------|
| #INL (occ.) | 354 (959) | 307 (1151) | 208 (2071) | 301 (1082) | 236 (1521) |
| #OOL (occ.) | 20 (55) | 91 (351) | 15 (50) | 103 (162) | 16 (39) |
| #SING (occ.) | 340 (984) | 380 (1280) | 198 (2093) | 383 (1186) | 252 (1560) |
| #MULTI (occ.) | 34 (30) | 18 (222) | 25 (28) | 21 (58) | 0 (0) |
| #INV (occ.) | 292 (668) | 312 (1263) | 192 (1749) | 316 (1035) | 204 (1375) |
| #OOV (occ.) | 82 (346) | 86 (239) | 31 (372) | 88 (209) | 48 (185) |

Regarding the QbE STD task, three different acoustic examples per query were provided for both development and test datasets. One example was extracted from the same dataset as the one to be searched (hence in-domain acoustic examples). This scenario considered the case in which the user finds a term of interest within a certain speech dataset and he/she wants to search for new occurrences of the same query.

The two other examples were recorded by the evaluation organizers and comprised an scenario where the user pronounces the query to be searched (hence, out-of-domain acoustic examples). These two out-of-domain acoustic examples amount to 3 s of speech

with PCM, 16 kHz, single channel and 16 bits per sample with the microphone of an HP ProBook Core i5, 7th Gen and with a Sennheiser SC630 USB CTRL microphone with noise cancellation, respectively.

The queries employed for the QbE STD task were chosen from the STD queries, and the corresponding figures are presented in Table 5. For both the STD and QbE STD tasks, a multi-word query was considered OOV in the case where any of the words that form the query were OOV.

Table 5. Development and test query list characteristics for the MAVIR, RTVE and SPARL20 databases for the QbE STD task. ‘dev’ stands for development, ‘INL’ refers to in-language queries, ‘OOL’ to foreign terms, ‘SING’ to single-word queries, ‘MULTI’ to multi-word queries, ‘INV’ to in-vocabulary queries, ‘OOV’ to out-of-vocabulary queries and ‘occ.’ stands for occurrences.

| Query List | Dev-MAVIR | Dev-RTVE | Test-MAVIR | Test-RTVE | Test-SPARL20 |
|---------------|-----------|-----------|------------|-----------|--------------|
| #INL (occ.) | 96 (386) | 81 (464) | 99 (1163) | 89 (808) | 87 (903) |
| #OOL (occ.) | 6 (39) | 22 (110) | 7 (29) | 19 (72) | 13 (30) |
| #SING (occ.) | 93 (407) | 101 (544) | 100 (1180) | 105 (861) | 100 (933) |
| #MULTI (occ.) | 9 (18) | 2 (30) | 6 (12) | 3 (19) | 0 (0) |
| #INV (occ.) | 83 (296) | 76 (480) | 94 (979) | 87 (750) | 65 (788) |
| #OOV (occ.) | 19 (129) | 27 (94) | 12 (213) | 21 (130) | 35 (145) |

2.3. Evaluation Metrics

In search on speech systems (both for STD and QbE STD tasks), a hypothesised occurrence is called a *detection*; if the detection corresponds to an actual occurrence, it is called a *hit*, otherwise it is called a *false alarm*. If an actual occurrence is not detected, this is called a *miss*. The actual term weighted value (ATWV) metric proposed by NIST [138] was used as the main metric for the evaluation. This metric combines the hit rate and false alarm rate of each query and averages over all the queries, as shown in Equation (1):

$$ATWV = \frac{1}{|\Delta|} \sum_{Q \in \Delta} \left(\frac{N_{hit}^Q}{N_{true}^Q} - \beta \frac{N_{FA}^Q}{T - N_{true}^Q} \right), \quad (1)$$

where Δ denotes the set of queries and $|\Delta|$ is the number of queries in this set. N_{hit}^Q and N_{FA}^Q represent the numbers of hits and false alarms of query Q , respectively, and N_{true}^Q is the number of actual occurrences of query Q in the audio. T denotes the audio length in seconds, and β is a weight factor set to 999.9, as in the ATWV proposed by NIST [37]. This weight factor causes an emphasis placed on recall compared to precision with a ratio 10:1.

The time tolerance for query detection was higher than the original proposed by NIST to encourage participants to build end-to-end systems for both STD and QbE STD tasks. To do so, a detection was labelled as correct in case it appeared within ± 15 -s interval from the ground-truth timestamp.

ATWV represents the term weighted value (TWV) for an *optimal* threshold given by the system (usually tuned on the development data). An additional metric, called maximum term weighted value (MTWV) [138] is also used in this paper to evaluate the upper-bound system performance regardless of the decision threshold.

Additionally, $p(\text{Miss})$ and $p(\text{FA})$ values, which represent the probability of miss and FA of the system as defined in Equations (2) and (3), respectively, are also reported.

$$p(\text{Miss}) = 1 - \frac{N_{hit}}{N_{true}} \quad (2)$$

$$p(\text{FA}) = \frac{N_{\text{FA}}}{T - N_{\text{true}}}, \quad (3)$$

where N_{hit} is the number of hits obtained by the system, N_{true} is the actual number of occurrences of the queries in the audio, N_{FA} is the number of FAs produced by the system and T denotes the audio length (in seconds). These values, therefore, provide a quantitative way to measure system performance in terms of misses (or equivalently, hits) and false alarms.

In addition to ATWV, MTWV, p(Miss) and p(FA) figures, NIST also proposed a detection error tradeoff (DET) curve [141] that evaluates the performance of a system at various miss/FA ratios. Although DET curves were not used for the evaluation itself, they are also presented in this paper for system comparison.

The NIST STD evaluation tool [142] was employed to compute both the MTWV, ATWV, p(Miss) and p(FA) figures, along with the DET curves.

2.4. Comparison with Previous Search on Speech International Evaluations

The SoS ALBAYZIN evaluation comprises two different tasks (STD and QbE STD). The most similar evaluations to the SoS ALBAYZIN evaluation are the NTCIR-11 [143] and NTCIR-12 [144] search on speech evaluations that also involved these two tasks. The data used in these NTCIR evaluations contained spontaneous speech in Japanese provided by the National Institute for Japanese Language and spontaneous speech recorded during seven editions of the Spoken Document Processing Workshop.

These evaluations also provided the participants a voice activity detector (VAD) for the speech data, the manual transcription of the speech data and the output of an LVCSR system. The results of those evaluations vary from 0.6140 to 0.7188 in terms of F-measure for the STD task and from 0.5860 to 0.7963 in terms of MAP measure for the QbE STD task.

However, our SoS evaluation differs from those in several aspects:

- The SoS ALBAYZIN evaluation makes use of a different language (i.e., Spanish).
- The SoS ALBAYZIN evaluation defines disjoint development and test query lists, along with different domains and an *unseen* domain for test data to measure the generalization capability of the systems.
- Participants were highly encouraged to build end-to-end systems.
- In the case that participants do not build end-to-end systems, the SoS ALBAYZIN evaluation defines two types of queries: INV and OOV, which demand participants to build different types of systems, especially those handling OOV query search.
- The SoS ALBAYZIN evaluation for the QbE STD task provides two acoustic query types: in-domain acoustic examples, which correspond to spoken queries extracted from the search speech collection; and out-of-domain acoustic examples, which correspond to spoken queries recorded by the evaluation organizers.

2.4.1. Comparison with Previous STD International Evaluations

Since the SoS ALBAYZIN evaluation integrates the STD task, it is worth mentioning other previous STD international evaluations. In 2006, NIST launched the first STD evaluation [145]. This evaluation involved different languages (i.e., English, Arabic and Mandarin) and different acoustic conditions (i.e., conversational telephone speech (CTS), broadcast news and round-table meetings). The best performance was obtained for the broadcast news condition and English language (ATWV = 0.8485), for which more data for system construction was typically available by that date.

IARPA BABEL program started in 2011 and addressed KWS/STD tasks to a great extent as well [146]. This program focused on building fully automatic and noise-robust speech recognition and search systems in a very limited amount of time (e.g., one week) and with limited amount of training data. The languages addressed in that program were low-resourced, such as Cantonese, Pashto, Tagalog, Turkish, Vietnamese, Swahili, Tamil and so on, and significant research has been carried out [13,61,147–159].

Moreover, NIST continued organising STD evaluations in the so-called open keyword search (OpenKWS) evaluations from 2013 to 2016 [160–163]. These OpenKWS evaluations were very similar to the NIST STD 2006 evaluation, since they included CTS data and microphone speech data. The main novelty was that the evaluation language was unknown up to 4 weeks (or less) before the system submission. The main results of these evaluations are presented in Table 6.

Table 6. Best performance (in terms of the Actual Term Weighted Value, ATWV) obtained in the different editions (2013, 2014, 2015 and 2016) of the OpenKWS evaluations under the full language pack condition.

| Evaluation | ATWV | Language |
|--------------|--------|------------|
| OpenKWS 2013 | 0.6248 | Vietnamese |
| OpenKWS 2014 | 0.5802 | Tamil |
| OpenKWS 2015 | 0.6548 | Swahili |
| OpenKWS 2016 | 0.8730 | Georgian |

From 2017, NIST has also been organising the biennial open speech analytics technologies (OpenSAT) evaluations [164–167], which include keyword search among their tasks, and aim to provide support for developing novel speech analytic technologies for low-resourced languages by including multiple speech analytic tasks (speech activity detection, speech recognition and keyword spotting) and multiple data domains (low-resourced languages, speech from video and public-safety communications). The only publicly available results are those from the 2017 evaluation and show an ATWV = 0.57 for the low-resourced languages and a quite low ATWV (negative value) for the challenging public-safety communication domain.

In the STD task of the SoS ALBAYZIN evaluation, speech comprises various recording conditions: (1) real talks in real workshops held in large conference rooms with audience, (2) broadcast news speech and (3) parliament sessions speech. In the workshop data, microphones, conference rooms and even recording conditions change from one recording to another, where tabletop and ground standing microphones were typically employed. In addition, the SoS ALBAYZIN evaluation explicitly defines different in-vocabulary and out-of-vocabulary term sets and makes use of Spanish language. These differences in the evaluation conditions make our evaluation pose different challenges but also make it difficult to compare the results obtained in our evaluation with those of previous NIST STD/OpenKWS/OpenSAT evaluations.

2.4.2. Comparison with Previous Qbe STD International Evaluations

Several QbE STD international evaluations held around the world are similar, in some ways, to the QbE STD task of the SoS ALBAYZIN evaluation. These comprise the spoken web search (SWS) task in MediaEval 2011 [168], 2012 [169] and 2013 [170] evaluations, whose results range from 0.173 to 0.762 in terms of ATWV.

However, the QbE STD task of the SoS ALBAYZIN evaluation differs from those evaluations in several aspects:

- The most important difference is the nature of the audio content. In the SWS evaluations, the speech is typically telephone speech, either conversational or read and elicited speech, or speech recorded with in-room microphones. In the SoS ALBAYZIN evaluation, the audio consists of microphone recordings of real talks in workshops that took place in large conference rooms in the presence of audience. Microphones, conference rooms and recording conditions change from one recording to another. The microphones were not close talking microphones but table top or floor standing microphones. In addition, the SoS ALBAYZIN evaluation also contains broadcast TV shows and live-talking parliament sessions speech, and explicitly defines different in-vocabulary and out-of-vocabulary query sets.

- SWS evaluations dealt with Indian and African-derived languages, as well as Albanian, Basque, Czech, non-native English, Romanian and Slovak languages, while the SoS ALBAYZIN evaluation only deals with the Spanish language.

This makes it difficult to compare the results obtained in the QbE STD task of the SoS ALBAYZIN evaluation with those of SWS MediaEval evaluations.

2.4.3. Comparison with Previous Search on Speech Albayzin Evaluations

From 2012, the SoS ALBAYZIN evaluation is being organized every two years. Compared with the previous evaluation held in 2018, the 2020 edition integrates the following advances:

- The Spanish parliament sessions is a new domain that was selected as the *unseen* domain to test the system generalization capability.
- For the QbE STD task, organizers recorded two acoustic examples per query aiming to encourage participants to build a more robust acoustic query for search.
- Aiming to build end-to-end systems both for STD and QbE STD tasks, organizers allowed participants to include OOV queries within the system dictionary in the case an end-to-end system is built.

3. Systems

Seven different systems were submitted from two research teams to the SoS evaluation. Specifically, four systems were submitted for the STD task, and three systems were submitted for the QbE STD task. This allows for a comparison between both tasks. Table 7 lists the systems submitted along with their main characteristics.

Table 7. Participants in the ALBAYZIN 2020 SoS along with the systems submitted.

| Team ID | Research Institution | Systems | Task | Type of System |
|---------|--|-----------------|---------|----------------|
| CENATAV | Voice group, Advanced Technologies Application Center, Cuba | Kaldi DNN + DA | STD | LVCSR |
| | | Kaldi SGMM + DA | STD | LVCSR |
| | | Kaldi DNN | STD | LVCSR |
| | | Kaldi SGMM | STD | LVCSR |
| AUDIAS | Universidad Autónoma de Madrid, Spain | E2E + LA | QbE STD | LD end-to-end |
| | | E2E | QbE STD | LI end-to-end |
| | | E2E + ZNORM | QbE STD | LI end-to-end |

3.1. Spoken Term Detection

The four systems submitted to the STD task are based on LVCSR from the Kaldi toolkit [74]. These systems are described below.

3.1.1. Kaldi-Based DNN with Data Augmentation System (Kaldi DNN + DA)

This system, whose architecture is presented in Figure 1, is based on an LVCSR system constructed with the open-source Kaldi toolkit [74].

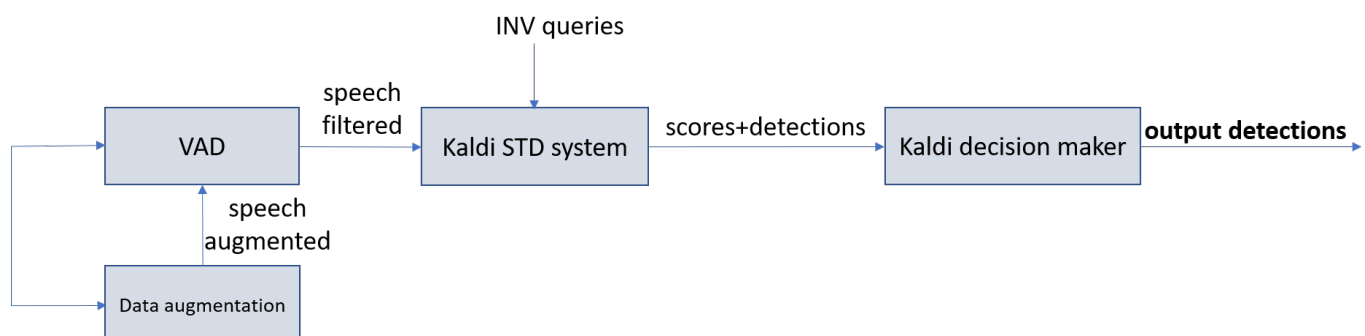


Figure 1. Architecture of the Kaldi DNN + DA system.

First, a VAD is employed to remove noisy and voiceless segments from the original speech signals. This VAD was developed with the open-source machine learning framework Pytorch as an artificial neural network that consists of a single unidirectional long short-term memory (LSTM) recurrent layer and one fully connected layer. The input features for this VAD are 12 Mel frequency cepstrum coefficients (MFCCs) along with their first derivatives. Data employed for VAD training consist of the training data provided by the evaluation organizers.

Then, the LVCSR module based on Kaldi takes the speech segments as input and produces word lattices as output. The system's design makes use of the s5 Wall Street Journal (WSJ) recipe in Kaldi [171]. The acoustic features are 13 MFCCs with cepstral mean and variance normalization (CMVN) applied to reduce the effects of the channel, and appended, just for the flat initialization of the acoustic model, with their delta and acceleration coefficients.

Some other transformations, such as linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT) and feature-space maximum likelihood linear regression (fMLLR) are also applied to obtain more robust features. The acoustic model training starts with a flat initialization of context-independent phone hidden Markov models (HMMs). Then, several re-training and alignment of acoustic models are carried out to produce context-dependent phone HMMs, following the standard procedures of the Kaldi s5 WSJ recipe.

These phone models consist of three HMM states, each in a tied-pdf cross-word tri-phone context with Gaussian mixture models (GMMs). Then, the subspace Gaussian mixture model (SGMM) is employed on top of the GMM-HMM model for speaker adaptation, as described in [172], using fMLLR features and sharing the same Gaussian model. The GMM-HMM is also employed to produce the alignments needed to the DNN-based acoustic model (DNN-HMM). The DNNs consist of two hidden layers with 300 nodes each. The number of spliced frames is 9 to produce 360 dimensional vectors as input to the first layer. The output layer is a soft-max layer representing the log-posteriors of the context-dependent HMM states.

The acoustic model training data comprise the following data: the TC-STAR evaluation package data recorded from 2005 to 2007, which amount to about 27 h of speech; a subset of the *dev1* set of the RTVE data, which amounts to about 14 h of speech; the LibriVox-Bailén data, which consist of audiobook speech and amount to about 7 h of speech; and the MAVIR training data, which amount to more than 4 h of speech. There are 52 h of speech material in total in the four datasets. Overlapped speech is removed from the *dev1* set of RTVE, and thus eventually 51 h of speech are used for acoustic model training.

The data used for language model training consist of the text transcriptions of the acoustic model training data, which contain 163,000 word occurrences. Specifically, these text transcriptions are given to the SRILM toolkit [173] to create a trigram-based language model with 45,000 trigrams, 188,000 bigrams and 28,000 unigrams. The system's vocabulary consists of the different words contained in the training data and, after removing the OOV words, amounts to 28,000 words. The multilingual G2P transcriber [174] is employed to obtain the phone transcription of each word.

Then, the Kaldi decoder outputs word lattices using the DNN-HMM based acoustic models. The STD subsystem, which takes the word lattices as input, includes the Kaldi term detector and Kaldi decision maker. The Kaldi term detector [74,76,77] searches for the input queries within the word lattices. To do so, these lattices are first processed using the lattice indexing technique described in [175] so that these are converted from individual weighted finite state transducers (WFSTs) to a single generalized factor transducer structure in which the start-time, end-time and lattice posterior probability of each word token are stored as 3-dimensional vectors.

This factor transducer represents an inverted index of all word sequences in the lattices. Thus, given a list of queries in a written form, a finite state machine is first created, then it accepts each query and finally composes it with the factor transducer to obtain all

occurrences of the queries in the search speech files. The Kaldi decision maker provides a YES/NO decision for each detection based on the term specific threshold (TST) approach presented in [54], so that a score for each detection is computed as shown in Equation (4):

$$p > \frac{N_{conf}}{\frac{T}{\beta} + \frac{\beta-1}{\beta} N_{conf}}, \quad (4)$$

where p is the confidence score of the detection, N_{conf} is the sum of the confidence score of all the detections of the given query, β is set to 999.9 (as in Equation (1)), and T is the length of the audio (in seconds).

No strategy was employed for retrieving the OOV words, and therefore this system cannot detect the OOV terms.

In order to alleviate the data mismatch between training and test speech, a data augmentation strategy is incorporated to the system. This aims to generate artificial (i.e., synthetic) speech data to augment the speech training material. Both noise databases and a tool to simulate noisy conditions in the speech with different signal-to-noise ratio (SNR) levels were employed.

The noise resources consist of two different databases: (1) free TV programs and radio music background excerpts collected through the Internet; and (2) the DEMAND database [176]. The noise effects in the DEMAND database are common in-door and out-door background sounds, which aim to mimic acoustic noises that occur in natural conversations, such as those in the MAVIR database. Both TV programs and radio music background excerpts aim to mimic the acoustic events that may occur in the RTVE database.

The filtering and noise adding tool (FaNT), which was originally used to create noisy data in the AURORA 2 speech recognition corpus, was employed to synthesize the augmented speech data [177]. This tool allows adding noise to previously recorded speech signals with the desired SNR. Two different SNR ranges were employed for the two noise databases (i.e., 1–10 dB and 6–15 dB) so that the training data are augmented by a factor of 4.

3.1.2. Kaldi-Based SGMM with Data Augmentation System (Kaldi SGMM + DA)

This system is the same as the Kaldi DNN + DA system described before, except that SGMMs are employed as acoustic models.

3.1.3. Kaldi-Based DNN System (Kaldi DNN)

This system is the same as the Kaldi DNN + DA system described before, except that the data augmentation strategy is not applied.

3.1.4. Kaldi-Based SGMM System (Kaldi SGMM)

This system is the same as the Kaldi SGMM + DA system, except that the data augmentation strategy is not applied.

3.2. Query-by-Example Spoken Term Detection

The three systems submitted to the QbE STD task are end-to-end systems based on deep learning approaches. These systems are described below.

3.2.1. End-to-End with Language Adaptation System (E2E + LA)

This system is based on attentive pooling networks [136], which were proposed in the context of question answering in NLP. These networks use a two-way attention mechanism able to compare a query (question) and a document (answer) of different lengths by focusing on the most relevant parts of both the query and the document. This approach was recently proposed for the QbE STD task [137]. The system developed is based on this work and has the structure depicted in Figure 2.

First, the WebRTC VAD is employed to remove silence and random noise effects that may appear at the beginning and end of the audio queries.

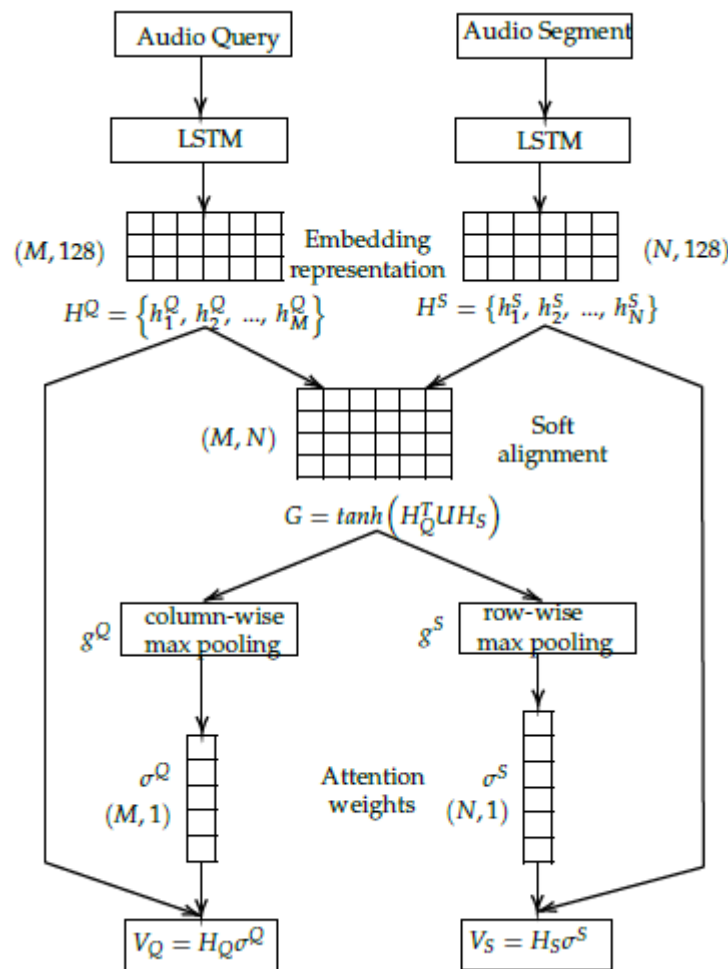


Figure 2. Attentive pooling network structure for the E2E + LA QbE STD system.

An LSTM network is used to obtain embedding representations of both the acoustic queries and the audio segments in which the utterances are divided. To do so, for a spoken query with acoustic features given by $Q = \{q_1, q_2, \dots, q_M\}$, where M is the number of frames of the query, the shared recurrent neural networks (RNNs) consisting of LSTM units project the query into a hidden state sequence $H_Q = \{h_1^Q, h_2^Q, \dots, h_M^Q\}$, where h_M^Q contains information of the whole query. In the same way, the audio segment with acoustic features given by $S = \{s_1, s_2, \dots, s_N\}$, where N is the number of frames of the audio segment, is encoded into the corresponding hidden state sequence $H_S = \{h_1^S, h_2^S, \dots, h_N^S\}$, hence the query and utterance representations convey the same feature space.

Then, the attention matrix, G , is computed as shown in Equation (5) so that the query and audio segment hidden state vector sequences can then be compared:

$$G = \tanh(H_Q^T U H_S). \tag{5}$$

Matrix U in Equation (5) is a measure of the H_Q and H_S representations and is learned during training. To build a more symmetric system, the measure matrix, U , is defined to be symmetric (i.e., $U = U^T$), which allows it to exchange query and audio segment representations, as given in Equation (6):

$$H_Q^T U H_S = H_S^T U^T H_Q = H_S^T U H_Q. \tag{6}$$

Matrix U is initialised with random normal samples with zero mean and 10^{-4} variance. Then, matrix G can be seen as a soft alignment score between each query and audio segment frame.

The next step is to apply column-wise and row-wise max-pooling to matrix G to generate the corresponding weight vectors $g^Q \in \mathbb{R}^M$ and $g^S \in \mathbb{R}^N$, which are computed as in Equations (7) and (8), respectively:

$$[g^Q]_j = \max_{1 \leq i \leq N} [G_{j,i}] \quad (7)$$

$$[g^S]_i = \max_{1 \leq j \leq M} [G_{j,i}]. \quad (8)$$

The j -th element in g^Q represents the weight applied to the j -th frame in the query Q . This results in an estimation of which frames of the queries are actually relevant for query/audio segment matching. Since g^Q and g^S are weight vectors, these are then normalized with a softmax function, whose output represents the attention weight vectors denoted as σ^Q and σ^S in Figure 2.

Finally, the dot product between the hidden states representing both the query and the audio segments in the utterance, and the attention vectors σ^Q and σ^S , is employed to enhance the meaningful information. To do so, V_Q and V_S are computed as in Equation (9):

$$V_Q = H_Q \sigma^Q, V_S = H_S \sigma^S. \quad (9)$$

The whole system is trained using a large margin cost function (i.e., hinge loss), whose aim is maximising the inter-class distance and minimising the intra-class distance. To do so, the training set comprises 3-element groups: a spoken query $Q = \{q_1, q_2, \dots, q_M\}$, a positive audio segment $S^{(P)} = \{s_1^{(P)}, s_2^{(P)}, \dots, s_N^{(P)}\}$ and a negative audio segment $S^{(N)} = \{s_1^{(N)}, s_2^{(N)}, \dots, s_{N'}^{(N)}\}$. The positive audio segment contains the same terms as the query, while the negative audio segment is a randomly-selected audio segment extracted from the training set that does not match with the query term.

Then, for each group, the tuples $(Q, S^{(P)})$ and $(Q, S^{(N)})$ are created to calculate the attention matrix, G , for each tuple and the corresponding vector representations, $(V_Q^{(P)}, V_S^{(P)})$ and $(V_Q^{(N)}, V_S^{(N)})$. The V_Q vector representation differs depending on the audio segment for which it is computed, due to the two-way attention mechanism. The cosine distance is employed to measure the similarity between the tuples V_Q and V_S , as in Equation (10):

$$l(V_Q, V_S) = (1 - \cos(V_Q, V_S))/2. \quad (10)$$

To minimise the distance between the query and the positive audio segment and maximise the distance between the query and the negative audio segment, the hinge objective function in Equation (11) is used.

$$L_{hinge} = \max\{0, M + l((V_Q^{(P)}, V_S^{(P)})) - l((V_Q^{(N)}, V_S^{(N)}))\}, \quad (11)$$

where M is the maximum allowable distance, which is set to 1.

For query and audio segment representation, 13-dimensional MFCC extracted using the Kaldi toolkit [74] and Python Speech Features library are employed. For MFCC computation, a 0.025 s-length sliding window with a shift equal to 0.01 s is used. The shared RNNs consist of two layers with 128 LSTM units for all the models. An Adam optimizer with a learning rate of 0.00005 and a minibatch of 128 are employed. The neural networks are implemented in Pytorch and are trained for four epochs.

The system was initially trained on the English LibriSpeech database [178]. Specifically, 500 randomly selected terms that have at least six phonemes and last between 0.5 and 1.0 s were employed to form training samples from 2 audio segments that contain the same term (one for the query (Q) and another for the positive audio segment (S^P)) and one audio segment that contains a different term for the negative audio segment (S^N). For each

training epoch, there were 1000 groups in total, which consisted of the (Q, S^P, S^N) tuple each. Then, the neural networks in the two-way attention mechanism are retrained for two epochs using the MAVIR development data. This aims to benefit from transfer learning by adapting the initial system trained on a language-independent setup to a small database in Spanish.

The search of each query in each utterance applies an adaptive sliding window in the utterances to form the audio segments for attentive pooling network-based query/audio segment matching. The length of the sliding window depends on the length of the query. Different window lengths (i.e., 50, 100, 150 and 200 frames) with a 50% overlap are employed so that the window length for searching each query is the upper limit that best approximates the query length (e.g., if the query has 37 frames, the window length is set to 50; if the query has 160 frames, the window length is set to 200 and so on). Once the final vectors V_Q and V_S representing the query and the audio segment are obtained according to Figure 2, the cosine distance between both vectors is taken as the final score for each detection.

3.2.2. End-to-End without Language Adaptation System (E2E)

This system is the same as the E2E + LA system except that the retraining with the Spanish database is not applied, hence aiming to build a language-independent QbE STD system.

3.2.3. End-to-End with Z-Score Normalization System (E2E + ZNORM)

This system is the same as the E2E system except that, in this case, a z-score normalization approach was included. To do so, each detection score is normalised according to Equation (12):

$$z = \frac{x - \mu}{\sigma}, \quad (12)$$

where x represents the original detection score, μ is the mean of all the detection scores, and σ is the standard deviation of all the detection scores.

4. Results and Discussion

We present the results obtained by the systems submitted to the evaluation for both the STD and the QbE STD tasks, and both for the development and test data. In addition, we also compare the performance of the QbE STD systems submitted to the evaluation with a text-based STD system.

4.1. Spoken Term Detection

4.1.1. Development Data

Evaluation system results for the development data are presented in Tables 8 and 9 for MAVIR and RTVE data, respectively. For MAVIR data, results show quite low performance, which suggests the challenging conditions of MAVIR data. For RTVE data, better results are obtained. The data augmentation technique results in better performance for the system that employs DNN for acoustic modelling both for MAVIR and RTVE data. This better performance is statistically significant for a paired t -test ($p < 0.001$) over the system without data augmentation.

This could be due to the more varied acoustic and background noise conditions used to train the DNN. However, when using SGMM for acoustic modelling, data augmentation gains are not so clear. This could be due to the fact that data augmentation works well with robust learning schema (e.g., neural network-based approaches) as DNNs. However, when less parameters are involved in the model, as in SGMM-based approaches, data augmentation is not so powerful. The better performance obtained with SGMM over the DNN model without data augmentation could be due to the limited training material used for DNN training, which makes SGMM outperform DNNs.

For MAVIR data, the best performance obtained by the *Kaldi DNN + DA* system is statistically significant ($p < 0.001$) for a paired t -test compared with the *Kaldi SGMM + DA*

and the *Kaldi DNN* systems and weak significant ($p < 0.03$) compared with the *Kaldi SGMM* system. For RTVE data, the best performance of the *Kaldi SGMM* system is statistically significant ($p < 0.001$) for a paired *t*-test compared with the *Kaldi DNN* and the *Kaldi DNN + DA* systems.

Table 8. System results of the STD task for MAVIR development data.

| System ID | MTWV | ATWV | p(FA) | p(Miss) |
|-----------------|--------|---------|---------|---------|
| Kaldi DNN + DA | 0.0018 | −0.0150 | 0.00000 | 0.997 |
| Kaldi SGMM + DA | 0.0000 | −0.4697 | 0.00000 | 1.000 |
| Kaldi DNN | 0.0000 | −0.4819 | 0.00000 | 1.000 |
| Kaldi SGMM | 0.0000 | −0.1713 | 0.00000 | 1.000 |

Table 9. System results of the STD task for RTVE development data.

| System ID | MTWV | ATWV | p(FA) | p(Miss) |
|-----------------|--------|--------|---------|---------|
| Kaldi DNN + DA | 0.1026 | 0.1018 | 0.00001 | 0.892 |
| Kaldi SGMM + DA | 0.2879 | 0.2853 | 0.00001 | 0.701 |
| Kaldi DNN | 0.0070 | 0.0046 | 0.00000 | 0.991 |
| Kaldi SGMM | 0.2886 | 0.2858 | 0.00002 | 0.692 |

The DET curves for MAVIR and RTVE data are presented in Figures 3 and 4, respectively. For MAVIR data, the *Kaldi DNN + DA* system performs the best for a low FA ratio, and the *Kaldi DNN* system performs the best for a low miss ratio. For RTVE data, the *Kaldi DNN + DA* system performs the best for a low FA ratio, and the *Kaldi SGMM* system performs the best for a low miss ratio.

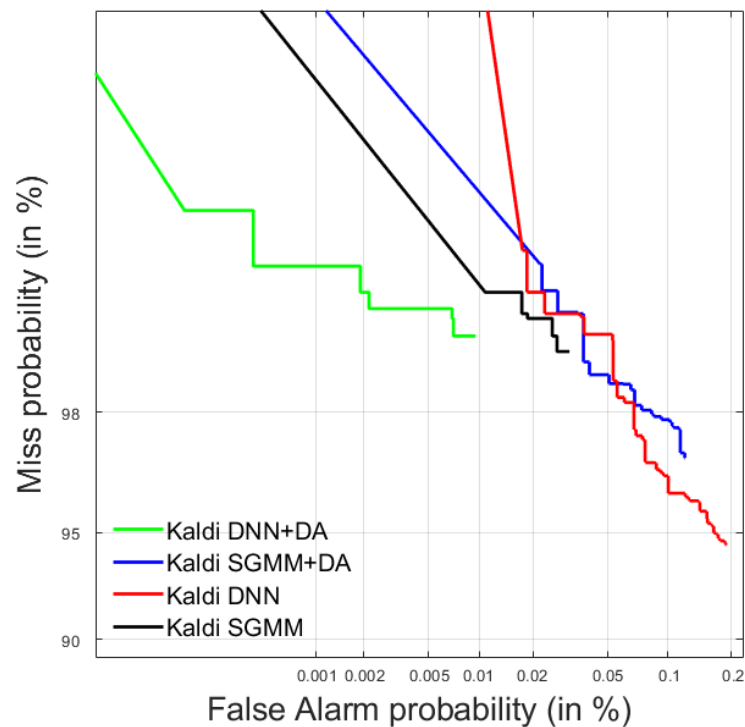


Figure 3. DET curves of the STD systems for MAVIR development data.

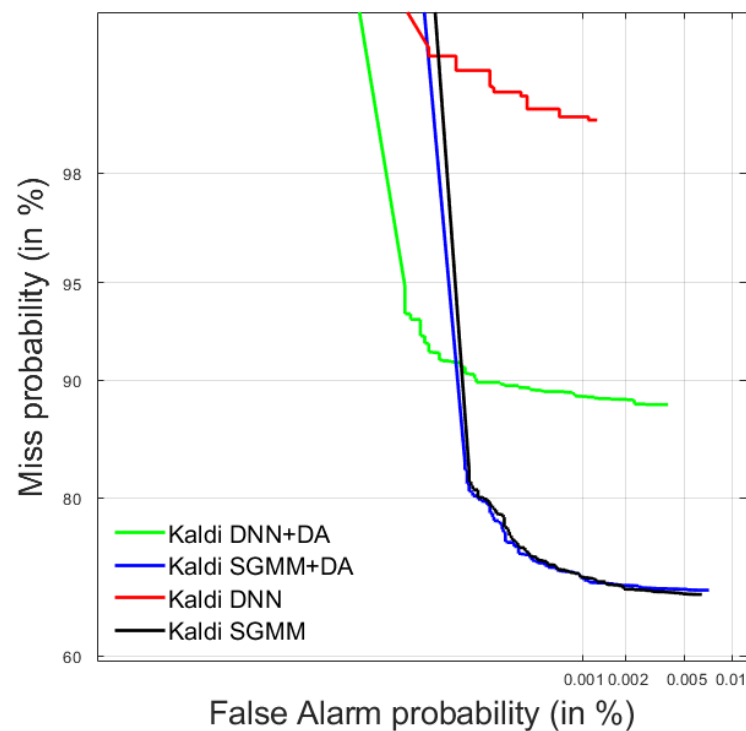


Figure 4. DET curves of the STD systems for RTVE development data.

4.1.2. Test Data

The evaluation system results on the test data are presented in Tables 10–12 for MAVIR, RTVE and SPARL20 data, respectively. Opposite to the results on the development data, the systems obtained better results for MAVIR data than for RTVE data. This could be due to the higher OOV rate for the RTVE data compared with for the MAVIR data (21.8% vs. 13.9%). Since the systems do not allow for OOV query detection, the performance drop is much more remarkable for RTVE data.

On the test data, the data augmentation technique does not improve the system performance neither for the DNN nor for the SGMM-based systems. This suggests some overfitting to the development data, where the DNN-based acoustic model benefits from that technique. The only exception is for the DNN acoustic model on the SPARL20 data, although the difference in performance is not statistically significant for a paired *t*-test. The results show that the best performance is, in general, obtained by the *Kaldi SGMM* system.

This best performance is statistically significant ($p < 0.001$) compared to the systems that employ data augmentation (i.e., *Kaldi DNN + DA* and *Kaldi SGMM + DA*) for MAVIR data and with the rest of the systems for SPARL20 data. For RTVE data, similar performance is obtained with the *Kaldi DNN* and the *Kaldi SGMM* systems, and the difference is not statistically significant.

To sum up, these results on the test data show that the data augmentation strategy fails when addressing spoken term detection on unseen test data.

Table 10. System results of the STD task for MAVIR test data.

| System ID | MTWV | ATWV | p(FA) | p(Miss) |
|-----------------|--------|---------|---------|---------|
| Kaldi DNN + DA | 0.0025 | −0.0168 | 0.00000 | 0.996 |
| Kaldi SGMM + DA | 0.0505 | 0.0403 | 0.00002 | 0.933 |
| Kaldi DNN | 0.4218 | 0.4230 | 0.00007 | 0.513 |
| Kaldi SGMM | 0.4413 | 0.4356 | 0.00007 | 0.489 |

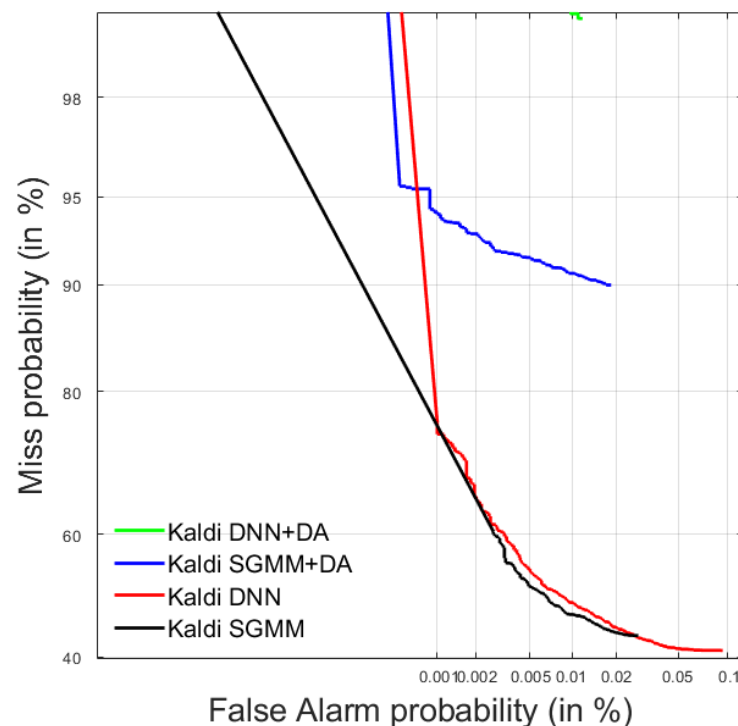
Table 11. System results of the STD task for RTVE test data.

| System ID | MTWV | ATWV | p(FA) | p(Miss) |
|-----------------|--------|---------|---------|---------|
| Kaldi DNN + DA | 0.0019 | −0.0049 | 0.00000 | 0.997 |
| Kaldi SGMM + DA | 0.0037 | −0.0010 | 0.00001 | 0.990 |
| Kaldi DNN | 0.2120 | 0.2123 | 0.00002 | 0.763 |
| Kaldi SGMM | 0.2107 | 0.2101 | 0.00001 | 0.778 |

Table 12. System results of the STD task for SPARL20 test data.

| System ID | MTWV | ATWV | p(FA) | p(Miss) |
|-----------------|--------|---------|---------|---------|
| Kaldi DNN + DA | 0.0096 | −0.0028 | 0.00001 | 0.985 |
| Kaldi SGMM + DA | 0.0149 | 0.0094 | 0.00000 | 0.980 |
| Kaldi DNN | 0.0074 | −0.0034 | 0.00000 | 0.989 |
| Kaldi SGMM | 0.5118 | 0.5090 | 0.00002 | 0.463 |

The DET curves for MAVIR, RTVE and SPARL20 data are presented in Figures 5–7, respectively. For MAVIR data, the *Kaldi SGMM* system performs the best for a low FA ratio, and the *Kaldi DNN* system performs the best for a low miss ratio. For RTVE data, the *Kaldi SGMM + DA* system performs the best for a low FA ratio, and the *Kaldi DNN* system performs the best for a low miss ratio. For SPARL20 data, the *Kaldi SGMM* system performs the best for both low FA and miss ratios.

**Figure 5.** DET curves of the STD systems for MAVIR test data.

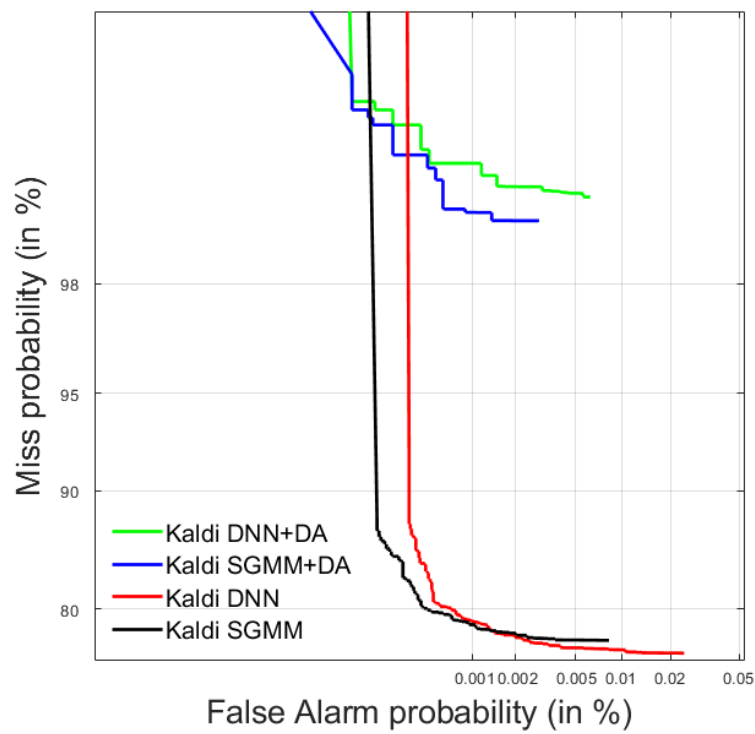


Figure 6. DET curves of the STD systems for RTVE test data.

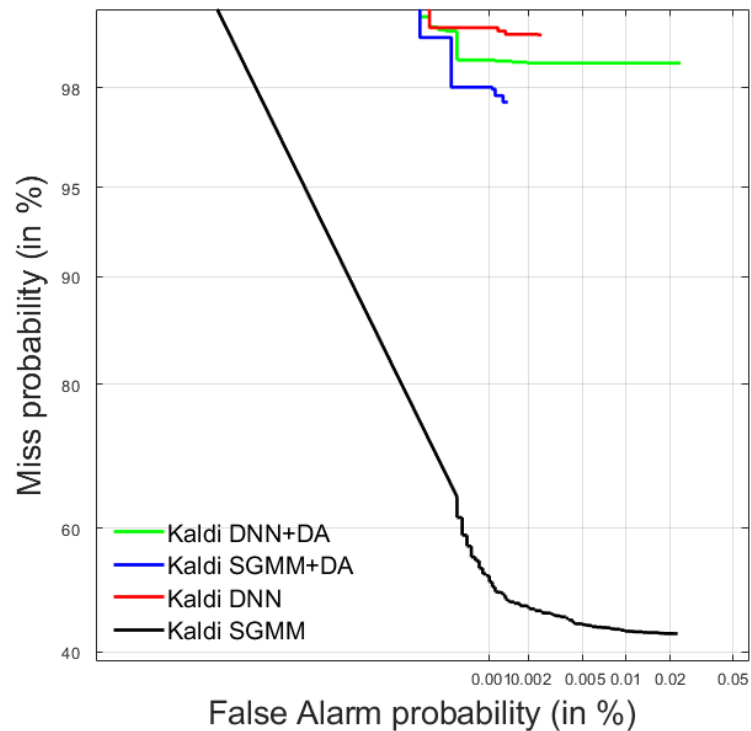


Figure 7. DET curves of the STD systems for SPARL20 test data.

4.2. Query-by-Example Spoken Term Detection

4.2.1. Development Data

Evaluation system results for the development data are presented in Tables 13 and 14 for MAVIR and RTVE data, respectively. They show similar findings: (1) The *E2E + LA* system obtains the best QbE STD performance for both datasets, since this system is adapted to the target language. This improvement is statistically significant for a paired

t -test ($p < 0.001$) for RTVE data and weak significant ($p < 0.03$) for MAVIR data compared to the other QbE STD systems. Moreover, this improvement is larger for MAVIR data (see MTWV results), since the language adaptation relies on these data; (2) The z -norm normalization decreases system performance. We consider that this could be due to the use of a global mean score and therefore global standard deviation score for all occurrences, regardless of the search query.

Comparing the QbE STD systems to a text-based STD system (i.e., the *Kaldi SGMM* system in the STD task) evaluated on the QbE STD queries, the text-based STD system obtains better performance for RTVE data, due to the use of the target language for system construction. This improvement is statistically significant for a paired t -test ($p < 0.001$) compared to the QbE STD systems.

Considering the results obtained by the end-to-end systems, it can be said that in general the performance is low. Preliminary experiments with the same systems were carried out on the read speech LibriSpeech dataset [178], following a similar experimental setup as in [137]. In this setup, the test language matched the training language and the query and each of the test word alignments were known in advance, so that words were only compared one to one (not one query against a segment of audio containing a few words, noise or fragments of words).

The results on this dataset with this setup showed a much better performance for P@20 (precision at 20) metric, similar to that presented in [137]. However, when facing the more challenging data conditions of this evaluation datasets, for which the speech scenarios are more complex, the test language differs from the training language and the query and test word alignments are both unknown, the performance drops dramatically.

Table 13. System results of the QbE STD task for MAVIR development data.

| System ID | MTWV | ATWV | p(FA) | p(Miss) |
|-------------|--------|-----------|---------|---------|
| E2E + LA | 0.0533 | 0.0491 | 0.00000 | 0.947 |
| E2E | 0.0160 | −38.5775 | 0.00000 | 0.984 |
| E2E + ZNORM | 0.0000 | −158.2873 | 0.00000 | 1.000 |
| Text STD | 0.0000 | −0.0508 | 0.00000 | 1.000 |

Table 14. System results of the QbE STD task for RTVE development data.

| System ID | MTWV | ATWV | p(FA) | p(Miss) |
|-------------|--------|----------|---------|---------|
| E2E + LA | 0.0465 | 0.0465 | 0.00000 | 0.954 |
| E2E | 0.0414 | −76.0473 | 0.00000 | 0.954 |
| E2E + ZNORM | 0.0000 | −51.9993 | 0.00000 | 1.000 |
| Text STD | 0.3101 | 0.3086 | 0.00001 | 0.677 |

The DET curves for MAVIR and RTVE data are shown in Figures 8 and 9, respectively. For MAVIR data, they show that the *E2E + LA* and the *E2E* systems perform the best for a low FA ratio, and the *E2E + ZNORM* system performs the best for a low miss ratio. For RTVE data, the *E2E + LA* system performs the best for a low FA ratio, whereas the *E2E + ZNORM* system performs the best for a low miss ratio.

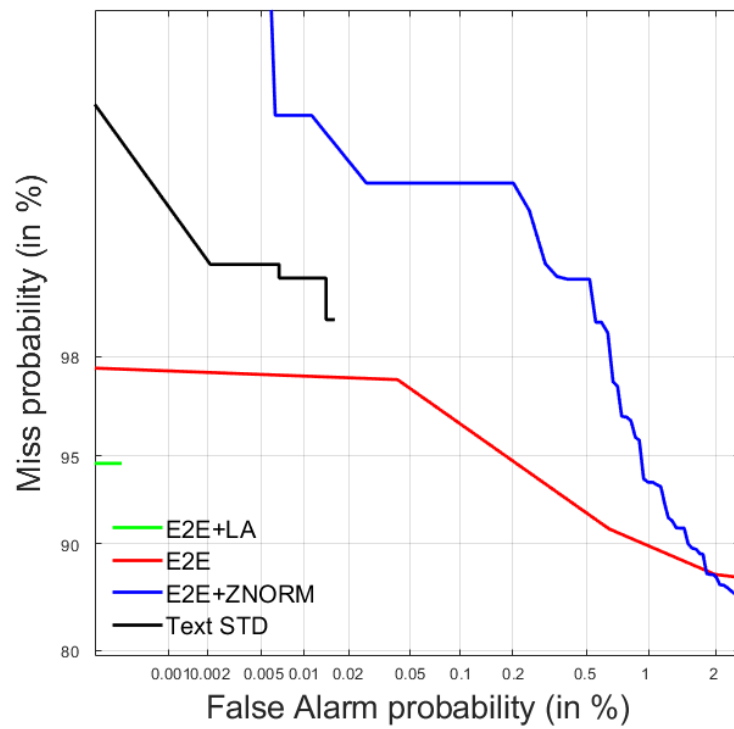


Figure 8. DET curves of the QbE STD systems and the text-based STD system for MAVIR development data.

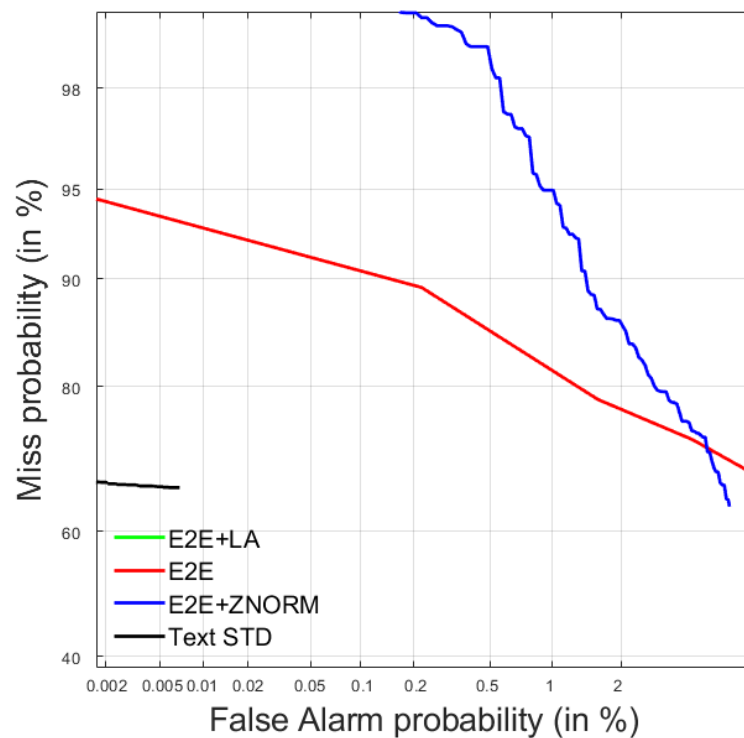


Figure 9. DET curves of the QbE STD systems and the text-based STD system for RTVE development data.

4.2.2. Test Data

Evaluation system results for the test data are presented in Tables 15–17 for MAVIR, RTVE and SPARL20 data, respectively. The performance for the QbE STD systems show, in general, the same findings as the development data. The performance obtained by the

end-to-end systems is low, opposite to the experiments carried out on the LibriSpeech dataset. For MAVIR and SPARL20 data, the best performance is obtained by the system that employs target language adaptation (i.e., the *E2E + LA* system).

This best performance is statistically significant for a paired *t*-test ($p < 0.001$) compared to the other QbE STD systems. For RTVE data, the best performance is for the *E2E + ZNORM* system, which does not employ language adaptation. This best performance is statistically significant for a paired *t*-test ($p < 0.001$). However, the worse MTWV performance obtained in comparison with the *E2E + LA* system for these RTVE data suggests some threshold calibration issues. This is confirmed by the results obtained by most of the systems, whose ATWV performance is below 0, even though the MTWV performance is above 0.

The text-based STD system (i.e., the *Kaldi SGMM* system in the STD task) evaluated on the QbE STD queries significantly outperforms the QbE STD systems ($p < 0.001$) for all the datasets. This suggests that the end-to-end systems for QbE STD need more research to approximate their performance to that of text-based STD. In particular, more research seems to be needed to adapt the system to the more realistic conditions of the evaluation, where the query and test word alignments are unknown.

Table 15. System results of the QbE STD task for MAVIR test data.

| System ID | MTWV | ATWV | p(FA) | p(Miss) |
|-------------|--------|-----------|---------|---------|
| E2E + LA | 0.0126 | −0.1061 | 0.00000 | 0.987 |
| E2E | 0.0000 | −393.5610 | 0.00000 | 1.000 |
| E2E + ZNORM | 0.0000 | −38.5959 | 0.00000 | 1.000 |
| Text STD | 0.4734 | 0.4682 | 0.00006 | 0.466 |

Table 16. System results of the QbE STD task for RTVE test data.

| System ID | MTWV | ATWV | p(FA) | p(Miss) |
|-------------|--------|-----------|---------|---------|
| E2E + LA | 0.0209 | −115.7086 | 0.00000 | 0.978 |
| E2E | 0.0209 | −88.3716 | 0.00000 | 0.978 |
| E2E + ZNORM | 0.0000 | −16.5831 | 0.00000 | 1.000 |
| Text STD | 0.3427 | 0.3413 | 0.00002 | 0.639 |

Table 17. System results of the QbE STD task for SPARL20 test data.

| System ID | MTWV | ATWV | p(FA) | p(Miss) |
|-------------|--------|-----------|---------|---------|
| E2E + LA | 0.0107 | 0.0107 | 0.00000 | 0.989 |
| E2E | 0.0306 | −34.2099 | 0.00000 | 0.961 |
| E2E + ZNORM | 0.0000 | −103.6805 | 0.00000 | 1.000 |
| Text STD | 0.3662 | 0.3583 | 0.00005 | 0.588 |

The DET curves are shown in Figures 10–12 for MAVIR, RTVE and SPARL20 data, respectively. For MAVIR data, they show that the best performance for a low FA ratio is for the *E2E + LA* system, whereas the text-based STD system performs the best for a low miss ratio. For RTVE data, the best performance for a low FA ratio is for the *E2E*, and the *E2E + LA* systems, and the text-based STD system performs the best for a low miss ratio. For SPARL20 data, the *E2E + LA* system performs the best for a low FA ratio, and the text-based STD system performs the best for a low miss ratio.

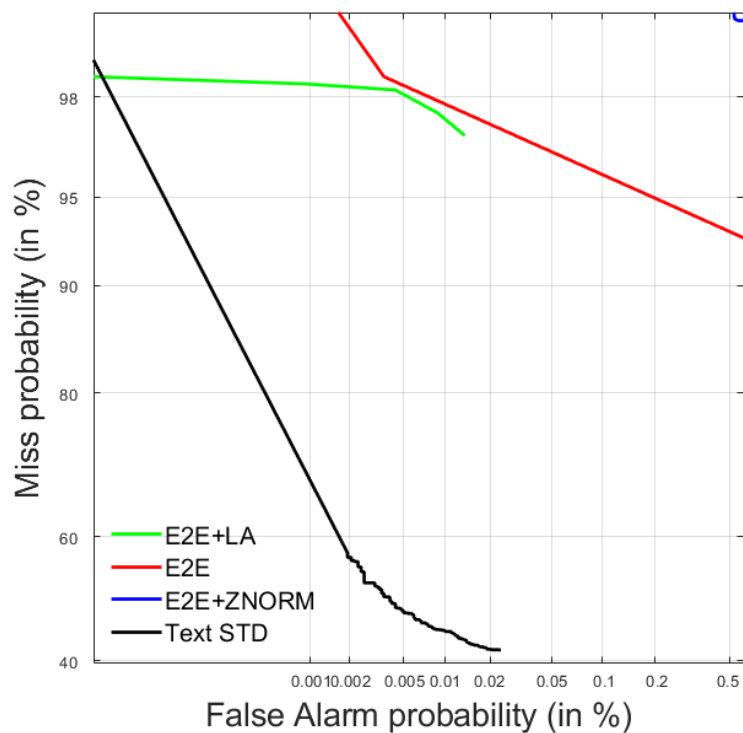


Figure 10. DET curves of the QbE STD systems and the text-based STD system for MAVIR test data.

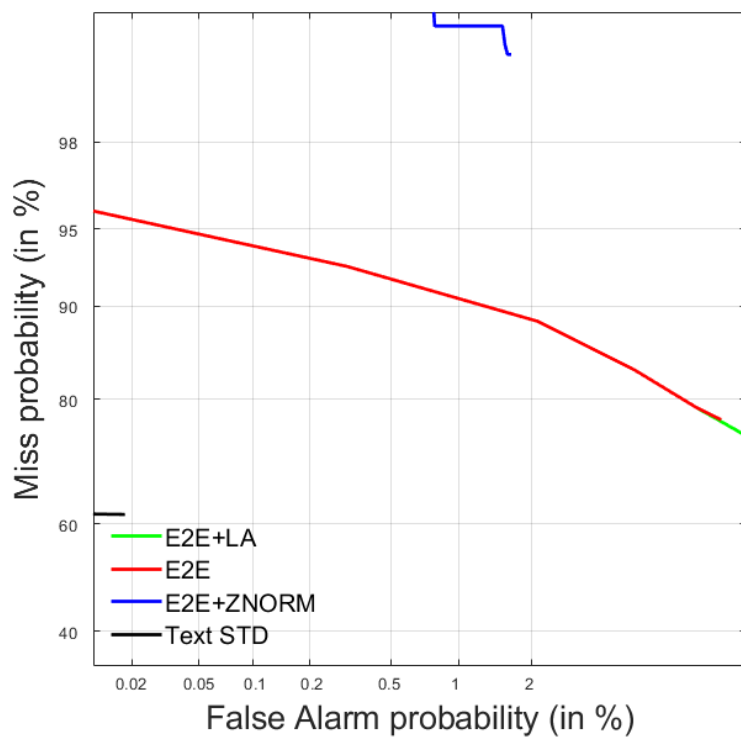


Figure 11. DET curves of the QbE STD systems and the text-based STD system for RTVE test data.

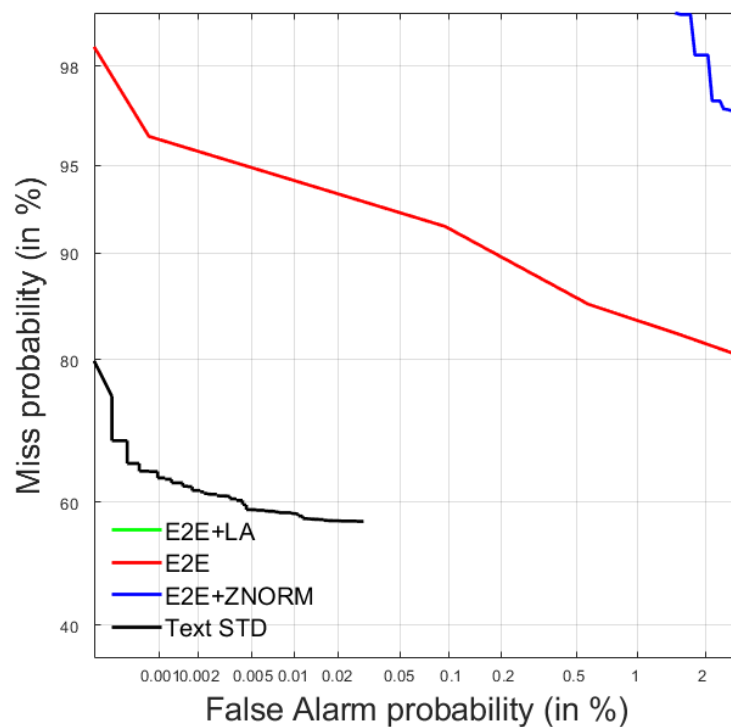


Figure 12. DET curves of the QbE STD systems and the text-based STD system for SPARL20 test data.

5. Post-Evaluation Analysis

A post-evaluation analysis for STD and QbE STD tasks was carried out in order to provide meaningful insights from diverse query properties.

5.1. Spoken Term Detection

Post-evaluation analyses for the STD task were performed based on INL and OOL query comparison and single-word and multi-word query comparison. Since the STD systems cannot retrieve OOV terms, the INV and OOV term comparison cannot be performed.

5.1.1. System Analysis for In-Language and Out-of-Language Queries

An additional analysis for INL and OOL queries was carried out for the STD task, and the results are presented in Tables 18–20 for MAVIR, RTVE and SPARL20 data, respectively. The results show in general better system performance for INL than for OOL queries (regardless of threshold calibration issues), since the target language matches that used for the system construction. The systems that do not employ data augmentation generally perform better than their counterparts with data augmentation as happened in the overall evaluation results.

Table 18. System results of the STD task for MAVIR test data for in-language (INL) and out-of-language (OOL) queries.

| System ID | INL | | OOL | |
|-----------------|--------|---------|--------|---------|
| | MTWV | ATWV | MTWV | ATWV |
| Kaldi DNN + DA | 0.0027 | −0.0154 | 0.0000 | −0.0367 |
| Kaldi SGMM + DA | 0.0517 | 0.0402 | 0.0408 | 0.0408 |
| Kaldi DNN | 0.4449 | 0.4462 | 0.1167 | 0.1075 |
| Kaldi SGMM | 0.4659 | 0.4598 | 0.1167 | 0.1075 |

Table 19. System results of the STD task for RTVE test data for in-language (INL) and out-of-language (OOL) queries.

| System ID | INL | | OOL | |
|-----------------|--------|---------|--------|---------|
| | MTWV | ATWV | MTWV | ATWV |
| Kaldi DNN + DA | 0.0025 | −0.0029 | 0.0000 | −0.0109 |
| Kaldi SGMM + DA | 0.0062 | 0.0012 | 0.0000 | −0.0075 |
| Kaldi DNN | 0.2633 | 0.2625 | 0.0727 | 0.0637 |
| Kaldi SGMM | 0.2584 | 0.2572 | 0.0716 | 0.0708 |

Table 20. System results of the STD task for SPARL20 test data for in-language (INL) and out-of-language (OOL) queries.

| System ID | INL | | OOL | |
|-----------------|--------|---------|--------|---------|
| | MTWV | ATWV | MTWV | ATWV |
| Kaldi DNN + DA | 0.0103 | −0.0024 | 0.0000 | −0.0087 |
| Kaldi SGMM + DA | 0.0160 | 0.0100 | 0.0000 | 0.0000 |
| Kaldi DNN | 0.0079 | −0.0036 | 0.0000 | 0.0000 |
| Kaldi SGMM | 0.5424 | 0.5395 | 0.0625 | 0.0625 |

5.1.2. System Analysis for Single and Multi-Word Queries

An additional analysis based on single and multi-word queries was carried out for the STD task, and the results are presented in Tables 21 and 22 for MAVIR and RTVE data, respectively. The results show, in general, better performance for single than multi-word queries, regardless of threshold calibration issues. The only exception is the case of the *Kaldi DNN* and the *Kaldi SGMM* systems for RTVE data, where these systems perform better on multi-word queries than on single-word queries.

Table 21. System results of the STD task for MAVIR test data for single-word (SING) and multi-word (MULTI) queries.

| System ID | SING | | MULTI | |
|-----------------|--------|---------|--------|--------|
| | MTWV | ATWV | MTWV | ATWV |
| Kaldi DNN + DA | 0.0028 | −0.0186 | 0.0000 | 0.0000 |
| Kaldi SGMM + DA | 0.0558 | 0.0445 | 0.0000 | 0.0000 |
| Kaldi DNN | 0.4495 | 0.4509 | 0.1839 | 0.1601 |
| Kaldi SGMM | 0.4705 | 0.4642 | 0.1667 | 0.1667 |

Table 22. System results of the STD task for RTVE test data for single-word (SING) and multi-word (MULTI) queries.

| System ID | SING | | MULTI | |
|-----------------|--------|---------|--------|--------|
| | MTWV | ATWV | MTWV | ATWV |
| Kaldi DNN + DA | 0.0020 | −0.0052 | 0.0000 | 0.0000 |
| Kaldi SGMM + DA | 0.0039 | −0.0011 | 0.0000 | 0.0000 |
| Kaldi DNN | 0.2063 | 0.2063 | 0.3161 | 0.3161 |
| Kaldi SGMM | 0.2048 | 0.2042 | 0.3161 | 0.3161 |

This discrepancy could be due to the fact that multi-word queries, on the one hand, may be easier to detect since they are typically longer than single-word queries; however, on the other hand, multi-word queries do convey more complex language modelling. Depending on the database domain, one factor can affect the system performance to a greater extent than another.

5.2. Query-by-Example Spoken Term Detection

Post-evaluation analyses for the QbE STD task rely on INL and OOL query comparison, INV and OOV query comparison and single-word and multi-word query comparison.

5.2.1. System Analysis for In-Language and Out-of-Language Queries

An analysis based on INL and OOL queries was carried out for the QbE STD systems and results are presented in Tables 23–25 for MAVIR, RTVE and SPARL20 data, respectively. These results show that OOL query detection obtains, in general and particularly for the systems that do not use language adaptation, better results than INL query detection, except for the RTVE data. This could be due to the fact that the QbE STD systems are built with English language, which matches the language of many OOL queries, particularly in MAVIR and SPARL20 databases.

Table 23. System results of the QbE STD task for MAVIR test data for in-language (INL) and out-of-language (OOL) queries.

| System ID | INL | | OOL | |
|-------------|--------|-----------|--------|-----------|
| | MTWV | ATWV | MTWV | ATWV |
| E2E + LA | 0.0135 | −0.1105 | 0.0357 | −0.0429 |
| E2E | 0.0000 | −402.2910 | 0.0000 | −270.0931 |
| E2E + ZNORM | 0.0000 | −37.0409 | 0.0000 | −60.5879 |

Table 24. System results of the QbE STD task for RTVE test data for in-language (INL) and out-of-language (OOL) queries.

| System ID | INL | | OOL | |
|-------------|--------|-----------|--------|-----------|
| | MTWV | ATWV | MTWV | ATWV |
| E2E + LA | 0.0241 | −111.3045 | 0.0058 | −136.3382 |
| E2E | 0.0241 | −83.7885 | 0.0058 | −109.8395 |
| E2E + ZNORM | 0.0000 | −17.5733 | 0.0000 | −11.9447 |

Table 25. System results of the QbE STD task for SPARL20 test data for in-language (INL) and out-of-language (OOL) queries.

| System ID | INL | | OOL | |
|-------------|--------|-----------|--------|----------|
| | MTWV | ATWV | MTWV | ATWV |
| E2E + LA | 0.0123 | 0.0123 | 0.0000 | 0.0000 |
| E2E | 0.0237 | −33.7965 | 0.0769 | −36.9764 |
| E2E + ZNORM | 0.0000 | −107.7061 | 0.0000 | −76.7394 |

5.2.2. System Analysis for In-Vocabulary and Out-of-Vocabulary Queries

A similar analysis for the QbE STD systems was carried out for INV and OOV queries and results are presented in Tables 26–28 for MAVIR, RTVE and SPARL20 data, respectively. They show that INV query detection is *easier* than OOV query detection, since better system performance is obtained for INV queries, regardless of threshold calibration issues. This

could be due to the fact that OOV queries exhibit more diverse properties in terms of occurrence rate and target language among others, which makes OOV query detection more complex inherently.

Table 26. System results of the QbE STD task for MAVIR test data for in-vocabulary (INV) and out-of-vocabulary (OOV) queries.

| System ID | INL | | OOL | |
|-------------|--------|-----------|--------|-----------|
| | MTWV | ATWV | MTWV | ATWV |
| E2E + LA | 0.0142 | −0.1196 | 0.0000 | 0.0000 |
| E2E | 0.0000 | −378.6430 | 0.0000 | −510.4187 |
| E2E + ZNORM | 0.0000 | −39.2731 | 0.0000 | −33.2913 |

Table 27. System results of the QbE STD task for RTVE test data for in-vocabulary (INV) and out-of-vocabulary (OOV) queries.

| System ID | INL | | OOL | |
|-------------|--------|-----------|--------|----------|
| | MTWV | ATWV | MTWV | ATWV |
| E2E + LA | 0.0259 | −127.9138 | 0.0000 | −65.1443 |
| E2E | 0.0259 | −97.8744 | 0.0000 | −49.0025 |
| E2E + ZNORM | 0.0000 | −18.8408 | 0.0000 | −7.2299 |

Table 28. System results of the QbE STD task for SPARL20 test data for in-vocabulary (INV) and out-of-vocabulary (OOV) queries.

| System ID | INL | | OOL | |
|-------------|--------|-----------|--------|----------|
| | MTWV | ATWV | MTWV | ATWV |
| E2E + LA | 0.0164 | 0.0164 | 0.0000 | 0.0000 |
| E2E | 0.0388 | −32.6405 | 0.0246 | −37.1246 |
| E2E + ZNORM | 0.0000 | −106.8322 | 0.0000 | −97.8272 |

5.2.3. System Analysis for Single and Multi-Word Queries

An analysis for single and multi-word queries in the QbE STD task was carried out and results are presented in Tables 29 and 30 for MAVIR and RTVE data, respectively. They show better performance for multi-word queries than single-word queries, since multi-word queries are typically longer than single-word queries. This produces less FAs so that the final performance improves.

Table 29. System results of the QbE STD task for MAVIR test data for single-word (SING) and multi-word (MULTI) queries.

| System ID | SING | | MULTI | |
|-------------|--------|-----------|--------|-----------|
| | MTWV | ATWV | MTWV | ATWV |
| E2E + LA | 0.0033 | −0.1224 | 0.1667 | 0.1667 |
| E2E | 0.0000 | −382.1081 | 0.0833 | −584.4417 |
| E2E + ZNORM | 0.0000 | −38.6250 | 0.0000 | −38.1104 |

Table 30. System results of the QbE STD task for RTVE test data for single-word (SING) and multi-word (MULTI) queries.

| System ID | SING | | MULTI | |
|-------------|--------|-----------|--------|----------|
| | MTWV | ATWV | MTWV | ATWV |
| E2E + LA | 0.0214 | −116.7881 | 0.0000 | −77.9245 |
| E2E | 0.0214 | −89.5284 | 0.0000 | −47.8828 |
| E2E + ZNORM | 0.0000 | −16.9599 | 0.0000 | −3.3951 |

6. Conclusions

This paper presented a multi-domain and international SoS ALBAYZIN 2020 evaluation, which comprises two different tasks: spoken term detection and query-by-example spoken term detection. The STD systems were based on the Kaldi toolkit with SGMM and DNN acoustic models. In both cases, the systems were evaluated with and without including a data augmentation strategy. The QbE STD task was addressed using neural end-to-end systems, which constitutes the first attempt regarding end-to-end system construction in the SoS ALBAYZIN evaluation series.

The most important conclusion drawn from this evaluation is that multi-domain search on speech is still challenging since the results showed a large variability when there is a change in the domain. On the other hand, the challenging conditions of the highly spontaneous speech in MAVIR data decreased the system performance with respect to unseen domains (i.e., SPARL20 data) for both STD and QbE STD tasks.

Regarding query properties, it was shown that INL query detection was *easier* than OOL query detection for the STD task, since the query language matched the target language. However, for the QbE STD task, for which the systems were trained on a different language that sometimes matched that of the OOL queries, OOL query detection is comparable or even improved the system performance for INL queries.

QbE STD systems also showed better performance for INV query detection than for OOV query detection, due to the more diverse properties of OOV queries.

Regarding single-word and multi-word query detection for the STD task, the evaluation results showed that single-word query detection was, in general, *easier* than multi-word query detection, since less words are needed to be detected. However, for the QbE STD task, the opposite occurred. This could be due to differences in the system architecture (standard ASR system vs. end-to-end system), along with the fact that longer queries typically produce less FAs on template-matching QbE STD scenarios and also in end-to-end neural systems, such as the one presented in this evaluation.

Given the best performance for each speech domain, there is still plenty of room for improvement. Specifically, the performance of the STD systems degrades with the data augmentation technique. This means that more research is needed in that topic to boost the performance. On the other hand, end-to-end systems for the QbE STD task deserve more research to improve their performance, in particular with respect to the realistic setup without information about query or test word alignments.

This encourages us to maintain the SoS ALBAYZIN evaluation in the next years, focusing on multi-domain SoS and the applicability of this technology to new unseen challenging domains. This evaluation suffered from a large decrease in the number of participants, likely due to the adverse global situation related to the COVID-19 pandemic. Specifically, in the next few months, we will be launching the ALBAYZIN 2022 STD evaluation to be held in November 2022 within the IberSPEECH conference.

This new evaluation edition, for which we hope to substantially increase the number of participants, aims to provide new domains and more challenging data (i.e., more difficult search terms) and evaluation conditions (i.e., rank the submitted systems from weighting the system performance according to the most challenging domain).

Author Contributions: Conceptualization, J.T. and D.T.T.; methodology, J.T., D.T.T., J.M.R., A.R.M. and J.I.A.-T.; software, J.T., D.T.T., J.M.R., A.R.M. and J.I.A.-T.; validation, J.T. and D.T.T.; formal analysis, J.T. and D.T.T.; investigation, J.T., D.T.T., J.M.R., A.R.M. and J.I.A.-T.; resources, J.T. and D.T.T.; data curation, J.T. and D.T.T.; writing—original draft preparation, J.T. and D.T.T.; writing—review and editing, J.T., D.T.T., J.M.R., A.R.M. and J.I.A.-T.; visualization, J.T., D.T.T., J.M.R., A.R.M. and J.I.A.-T.; supervision, J.T. and D.T.T.; project administration, J.T. and D.T.T.; funding acquisition, J.T. and D.T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science, Innovation and Universities of Spain, grant number RTI2018-095324-B-I00, and project DSForSec (grant number RTI2018-098091-B-I00). The APC was funded by the project DSForSec (grant number RTI2018-098091-B-I00) from the Ministry of Science, Innovation and Universities of Spain.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: MAVIR data are publicly available through this link: <http://cartago.illf.uam.es/mavir/index.pl>, accessed on 13 september 2021. RTVE data are also available upon request in <http://catedrartve.unizar.es/rtvedatabase.html>, accessed on 13 september 2021.

Acknowledgments: Authors also thank to Alicia Lozano Díez for the support on English edition.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ng, K.; Zue, V.W. Subword-based approaches for spoken document retrieval. *Speech Commun.* **2000**, *32*, 157–186. [[CrossRef](#)]
2. Chen, B.; Chen, K.Y.; Chen, P.N.; Chen, Y.W. Spoken Document Retrieval With Unsupervised Query Modeling Techniques. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 2602–2612. [[CrossRef](#)]
3. Lo, T.H.; Chen, Y.W.; Chen, K.Y.; Wang, H.M.; Chen, B. Neural relevance-aware query modeling for spoken document retrieval. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 466–473.
4. Heeren, W.; de Jong, F.M.; van der Werff, L.B.; Huijbregts, M.; Ordelman, R.J. Evaluation of spoken document retrieval for historic speech collections. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, Marrakech, Morocco, 26 May–1 June 2008; pp. 2037–2041.
5. Pan, Y.C.; Lee, H.Y.; Lee, L.S. Interactive Spoken Document Retrieval With Suggested Key Terms Ranked by a Markov Decision Process. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 632–645. [[CrossRef](#)]
6. Chen, Y.W.; Chen, K.Y.; Wang, H.M.; Chen, B. Exploring the Use of Significant Words Language Modeling for Spoken Document Retrieval. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Stockholm, Sweden, 20–24 August 2017; pp. 2889–2893.
7. Gao, P.; Liang, J.; Ding, P.; Xu, B. A novel phone-state matrix based vocabulary-independent keyword spotting method for spontaneous speech. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing—ICASSP '07, Honolulu, HI, USA, 15–20 April 2007; pp. 425–428.
8. Zhang, B.; Schwartz, R.; Tsakalidis, S.; Nguyen, L.; Matsoukas, S. White Listing and Score Normalization for Keyword Spotting of Noisy Speech. In Proceedings of the ISCA's 13th Annual Conference, Portland, OR, USA, 9–13 September 2012; pp. 1832–1835.
9. Mandal, A.; van Hout, J.; Tam, Y.C.; Mitra, V.; Lei, Y.; Zheng, J.; Vergyri, D.; Ferrer, L.; Graciarena, M.; Kathol, A.; et al. Strategies for High Accuracy Keyword Detection in Noisy Channels. In Proceedings of the Interspeech, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013; pp. 15–19.
10. Ng, T.; Hsiao, R.; Zhang, L.; Karakos, D.; Mallidi, S.H.; Karafiat, M.; Vesely, K.; Szoke, I.; Zhang, B.; Nguyen, L.; et al. Progress in the BBN Keyword Search System for the DARPA RATS Program. In Proceedings of the Interspeech, 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; pp. 959–963.
11. Mitra, V.; van Hout, J.; Franco, H.; Vergyri, D.; Lei, Y.; Graciarena, M.; Tam, Y.C.; Zheng, J. Feature fusion for high-accuracy keyword spotting. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 7143–7147.
12. Panchapagesan, S.; Sun, M.; Khare, A.; Matsoukas, S.; Mandal, A.; Hoffmeister, B.; Vitaladevuni, S. Multi-task learning and Weighted Cross-entropy for DNN-based Keyword Spotting. In Proceedings of the Interspeech, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 760–764.
13. Zhao, Z.; Zhang, W.Q. End-to-End Keyword Search Based on Attention and Energy Scorer for Low Resource Languages. In Proceedings of the Interspeech, 21st Annual Conference of the International Speech Communication Association, Shanghai, China, 25–29 October 2020; pp. 2587–2591.

14. Mamou, J.; Ramabhadran, B.; Siohan, O. Vocabulary independent spoken term detection. In Proceedings of the SIGIR '07: 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, 23–27 July 2007; pp. 615–622.
15. Schneider, D.; Mertens, T.; Larson, M.; Kohler, J. Contextual Verification for Open Vocabulary Spoken Term Detection. In Proceedings of the Interspeech, 11th Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010; pp. 697–700.
16. Parada, C.; Sethy, A.; Dredze, M.; Jelinek, F. A Spoken Term Detection Framework for Recovering Out-of-Vocabulary Words Using the Web. In Proceedings of the Interspeech, 11th Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010; pp. 1269–1272.
17. Szöke, I.; Fapšo, M.; Burget, L.; Černocký, J. Hybrid word-subword decoding for spoken term detection. In Proceedings of the 31st Annual International ACM SIGIR Conference, Singapore, 20–24 July 2008; pp. 42–48.
18. Wang, Y.; Metze, F. An In-Depth Comparison of Keyword Specific Thresholding and Sum-to-One Score Normalization. In Proceedings of the Interspeech, 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; pp. 2474–2478.
19. Mangu, L.; Saon, G.; Picheny, M.; Kingsbury, B. Order-free spoken term detection. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, Brisbane, Australia, 19–24 April 2015; pp. 5331–5335.
20. Fuchs, T.S.; Segal, Y.; Keshet, J. CNN-Based Spoken Term Detection and Localization without Dynamic Programming. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, 6–11 June 2021; pp. 6853–6857.
21. Lin, J.; Kilgour, K.; Roblek, D.; Sharifi, M. Training Keyword Spotters with Limited and Synthesized Speech Data. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, 4–8 May 2020; pp. 7474–7478.
22. Wintrode, J.; Wilkes, J. Fast Lattice-Free Keyword Filtering for Accelerated Spoken Term Detection. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, 4–8 May 2020; pp. 7469–7473.
23. Wang, Y.H.; Lee, H.Y.; Lee, L.S. Segmental Audio Word2Vec: Representing Utterances as Sequences of Vectors with Applications in Spoken Term Detection. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, 15–20 April 2018; pp. 6269–6273.
24. Kaneko, D.; Konno, R.; Kojima, K.; Tanaka, K.; Wook Lee, S.; Itoh, Y. Constructing Acoustic Distances Between Subwords and States Obtained from a Deep Neural Network for Spoken Term Detection. In Proceedings of the Interspeech, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 2879–2883.
25. Svec, J.; Psutka, J.V.; Smidl, L.; Trmal, J. A Relevance Score Estimation for Spoken Term Detection Based on RNN-Generated Pronunciation Embeddings. In Proceedings of the Interspeech, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 2934–2938.
26. Buzo, A.; Cucu, H.; Burileanu, C. Speed@MediaEval 2014: Spoken Term Detection with Robust Multilingual Phone Recognition. In Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, 16–17 October 2014; pp. 721–722.
27. Konno, R.; Ouchi, K.; Obara, M.; Shimizu, Y.; Chiba, T.; Hirota, T.; Itoh, Y. An STD system using multiple STD results and multiple rescoring method for NTCIR-12 SpokenQuery&Doc task. In Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan, 7–10 June 2016; pp. 200–204.
28. Jarina, R.; Kuba, M.; Gubka, R.; Chmulik, M.; Paralic, M. UNIZA System for the Spoken Web Search Task at MediaEval 2013. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013; pp. 791–792.
29. Anguera, X.; Ferrarons, M. Memory Efficient Subsequence DTW for Query-by-Example Spoken Term Detection. In Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME), San Jose, CA, USA, 15–19 July 2013; pp. 1–6.
30. Lin, H.; Stupakov, A.; Bilmes, J. Spoken keyword spotting via multi-lattice alignment. In Proceedings of the Interspeech, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, 22–26 September 2008; pp. 2191–2194.
31. Chan, C.; Lee, L. Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping. In Proceedings of the Interspeech, 11th Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010; pp. 693–696.
32. Mamou, J.; Ramabhadran, B. Phonetic Query Expansion for Spoken Document Retrieval. In Proceedings of the Interspeech, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, 22–26 September 2008; pp. 2106–2109.
33. Can, D.; Cooper, E.; Sethy, A.; White, C.; Ramabhadran, B.; Saraclar, M. Effect of pronunciations on OOV queries in spoken term detection. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, Taipei, Taiwan, 19–24 April 2009; pp. 3957–3960.
34. Rosenberg, A.; Audhkhasi, K.; Sethy, A.; Ramabhadran, B.; Picheny, M. End-to-end speech recognition and keyword search on low-resource languages. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 May 2017; pp. 5280–5284.

35. Audhkhasi, K.; Rosenberg, A.; Sethy, A.; Ramabhadran, B.; Kingsbury, B. End-to-end ASR-free keyword search from speech. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 May 2017; pp. 4840–4844.
36. Audhkhasi, K.; Rosenberg, A.; Sethy, A.; Ramabhadran, B.; Kingsbury, B. End-to-End ASR-Free Keyword Search From Speech. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1351–1359. [[CrossRef](#)]
37. Fiscus, J.G.; Ajot, J.G.; Garofolo, J.S.; Doddington, G. Results of the 2006 Spoken Term Detection Evaluation. In Proceedings of the ACM SIGIR Conference, Amsterdam, The Netherlands, 23–27 July 2007; pp. 45–50.
38. Hartmann, W.; Zhang, L.; Barnes, K.; Hsiao, R.; Tsakalidis, S.; Schwartz, R. Comparison of Multiple System Combination Techniques for Keyword Spotting. In Proceedings of the Interspeech, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 1913–1917.
39. Alumae, T.; Karakos, D.; Hartmann, W.; Hsiao, R.; Zhang, L.; Nguyen, L.; Tsakalidis, S.; Schwartz, R. The 2016 BBN Georgian telephone speech keyword spotting system. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 May 2017; pp. 5755–5759.
40. Vergyri, D.; Stolcke, A.; Gadde, R.R.; Wang, W. The SRI 2006 Spoken Term Detection System. In Proceedings of the NIST Spoken Term Detection Evaluation workshop (STD'06), Gaithersburg, MD, USA, 14–15 December 2006; pp. 1–15.
41. Vergyri, D.; Shafran, I.; Stolcke, A.; Gadde, R.R.; Akbacak, M.; Roark, B.; Wang, W. The SRI/OGI 2006 Spoken Term Detection System. In Proceedings of the Interspeech, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 27–31 August 2007; pp. 2393–2396.
42. Akbacak, M.; Vergyri, D.; Stolcke, A. Open-vocabulary spoken term detection using grapheme-based hybrid recognition systems. In Proceedings of the 33rd International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, Las Vegas, NV, USA, 30 March–4 April 2008; pp. 5240–5243.
43. Szöke, I.; Fapšo, M.; Karafiát, M.; Burget, L.; Grézl, F.; Schwarz, P.; Glembek, O.; Matějka, P.; Kopecký, J.; Černocký, J. Spoken Term Detection System Based on Combination of LVCSR and Phonetic Search. In *Machine Learning for Multimodal Interaction*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 4892, pp. 237–247.
44. Szöke, I.; Burget, L.; Černocký, J.; Fapšo, M. Sub-word modeling of out of vocabulary words in spoken term detection. In Proceedings of the 2008 IEEE Spoken Language Technology Workshop, Goa, India, 15–19 December 2008; pp. 273–276.
45. Meng, S.; Yu, P.; Liu, J.; Seide, F. Fusing multiple systems into a compact lattice index for Chinese spoken term detection. In Proceedings of the 33rd International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, Las Vegas, NV, USA, 30 March–4 April 2008; pp. 4345–4348.
46. Thambiratnam, K.; Sridharan, S. Rapid yet accurate speech indexing using dynamic match lattice spotting. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 346–357. [[CrossRef](#)]
47. Wallace, R.; Vogt, R.; Baker, B.; Sridharan, S. Optimising figure of merit for phonetic spoken term detection. In Proceedings of the 35th International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, Dallas, TX, USA, 15–19 March 2008; pp. 5298–5301.
48. Jansen, A.; Church, K.; Hermansky, H. Towards Spoken Term Discovery At Scale With Zero Resources. In Proceedings of the Interspeech, 11th Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010; pp. 1676–1679.
49. Parada, C.; Sethy, A.; Ramabhadran, B. Balancing false alarms and hits in spoken term detection. In Proceedings of the 35th International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, Dallas, TX, USA, 15–19 March 2008; pp. 5286–5289.
50. Trmal, J.; Wiesner, M.; Peddinti, V.; Zhang, X.; Ghahremani, P.; Wang, Y.; Manohar, V.; Xu, H.; Povey, D.; Khudanpur, S. The Kaldi OpenKWS System: Improving Low Resource Keyword Search. In Proceedings of the Interspeech, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 3597–3601.
51. Chen, C.P.; Lee, H.Y.; Yeh, C.F.; Lee, L.S. Improved Spoken Term Detection by Feature Space Pseudo-Relevance Feedback. In Proceedings of the Interspeech, 11th Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010; pp. 1672–1675.
52. Motlicek, P.; Valente, F.; Garner, P. English Spoken Term Detection in Multilingual Recordings. In Proceedings of the Interspeech, 11th Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010; pp. 206–209.
53. Szöke, I.; Fapšo, M.; Karafiát, M.; Burget, L.; Grézl, F.; Schwarz, P.; Glembek, O.; Matějka, P.; Kontár, S.; Černocký, J. BUT System for NIST STD 2006-English. In Proceedings of the NIST Spoken Term Detection Evaluation workshop (STD'06), Gaithersburg, MD, USA, 14–15 December 2006; pp. 1–15.
54. Miller, D.R.H.; Kleber, M.; Kao, C.L.; Kimball, O.; Colthurst, T.; Lowe, S.A.; Schwartz, R.M.; Gish, H. Rapid and Accurate Spoken Term Detection. In Proceedings of the Interspeech, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 27–31 August 2007; pp. 314–317.
55. Li, H.; Han, J.; Zheng, T.; Zheng, G. A Novel Confidence Measure Based on Context Consistency for Spoken Term Detection. In Proceedings of the Interspeech, 13th Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012; pp. 2430–2433.

56. Chiu, J.; Rudnicky, A. Using Conversational Word Bursts in Spoken Term Detection. In Proceedings of the Interspeech, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013; pp. 2247–2251.
57. Ni, C.; Leung, C.C.; Wang, L.; Chen, N.F.; Ma, B. Efficient methods to train multilingual bottleneck feature extractors for low resource keyword search. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 May 2017; pp. 5650–5654.
58. Meng, Z.; Juang, B.H. Non-Uniform Boosted MCE Training of Deep Neural Networks for Keyword Spotting. In Proceedings of the Interspeech, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 770–774.
59. Meng, Z.; Juang, B.H. Non-Uniform MCE Training of Deep Long Short-Term Memory Recurrent Neural Networks for Keyword Spotting. In Proceedings of the Interspeech, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 3547–3551.
60. Lee, S.W.; Tanaka, K.; Itoh, Y. Combination of diverse subword units in spoken term detection. In Proceedings of the Interspeech, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp. 3685–3689.
61. van Heerden, C.; Karakos, D.; Narasimhan, K.; Davel, M.; Schwartz, R. Constructing sub-word units for spoken term detection. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 May 2017; pp. 5780–5784.
62. Pham, V.T.; Xu, H.; Xiao, X.; Chen, N.F.; Chng, E.S. Pruning Strategies for Partial Search in Spoken Term Detection. In Proceedings of the International Symposium on Information and Communication Technology, Nha Trang, Vietnam, 7–8 December 2017; pp. 114–119.
63. Wollmer, M.; Schuller, B.; Rigoll, G. Keyword spotting exploiting Long Short-Term Memory. *Speech Commun.* **2013**, *55*, 252–265. [[CrossRef](#)]
64. Tejedor, J.; Toledano, D.T.; Wang, D.; King, S.; Colás, J. Feature analysis for discriminative confidence estimation in spoken term detection. *Comput. Speech Lang.* **2014**, *28*, 1083–1114. [[CrossRef](#)]
65. Zhuang, Y.; Chang, X.; Qian, Y.; Yu, K. Unrestricted Vocabulary Keyword Spotting using LSTM-CTC. In Proceedings of the Interspeech, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 938–942.
66. Pandey, L.; Nathwani, K. LSTM based Attentive Fusion of Spectral and Prosodic Information for Keyword Spotting in Hindi Language. In Proceedings of the Interspeech, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018; pp. 112–116.
67. Lileikyte, R.; Fraga-Silva, T.; Lamel, L.; Gauvain, J.L.; Laurent, A.; Huang, G. Effective keyword search for low-resourced conversational speech. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 May 2017; pp. 5785–5789.
68. Parlak, S.; Saraçlar, M. Spoken term detection for Turkish broadcast news. In Proceedings of the 33rd International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, Las Vegas, NV, USA, 30 March–4 April 2008; pp. 5244–5247.
69. Pham, V.T.; Xu, H.; Xiao, X.; Chen, N.F.; Chng, E.S. Re-ranking spoken term detection with acoustic exemplars of keywords. *Speech Commun.* **2018**, *104*, 12–23. [[CrossRef](#)]
70. Ragni, A.; Saunders, D.; Zahemszky, P.; Vasilakes, J.; Gales, M.J.F.; Knill, K.M. Morph-to-word transduction for accurate and efficient automatic speech recognition and keyword search. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 May 2017; pp. 5770–5774.
71. Chen, X.; Ragnil, A.; Vasilakes, J.; Liu, X.; Knill, K.; Gales, M.J.F. Recurrent neural network language models for keyword search. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 May 2017; pp. 5775–5779.
72. Xu, D.; Metzger, F. Word-based Probabilistic Phonetic Retrieval for Low-resource Spoken Term Detection. In Proceedings of the Interspeech, 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; pp. 2774–2778.
73. Khokhlov, Y.; Medennikov, I.; Romanenko, A.; Mendeleev, V.; Korenevsky, M.; Prudnikov, A.; Tomashenko, N.; Zatzvornitsky, A. The STC Keyword Search System For OpenKWS 2016 Evaluation. In Proceedings of the Interspeech, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 3602–3606.
74. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The KALDI Speech Recognition Toolkit. In Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2011, Waikoloa, HI, USA, 11–15 December 2011.
75. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Soplin, N.E.Y.; Heymann, J.; Wiesner, M.; Chen, N.; et al. ESPnet: End-to-End Speech Processing Toolkit. In Proceedings of the Interspeech, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018; pp. 2207–2211.
76. Chen, G.; Khudanpur, S.; Povey, D.; Trmal, J.; Yarowsky, D.; Yilmaz, O. Quantifying the Value of Pronunciation Lexicons for Keyword Search in Low Resource Languages. In Proceedings of the 38th International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013, Vancouver, BC, Canada, 26–31 May 2013; pp. 8560–8564.

77. Pham, V.T.; Chen, N.F.; Sivasdas, S.; Xu, H.; Chen, I.F.; Ni, C.; Chng, E.S.; Li, H. System and keyword dependent fusion for spoken term detection. In Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, USA, 7–10 December 2014; pp. 430–435.
78. Chen, G.; Yilmaz, O.; Trmal, J.; Povey, D.; Khudanpur, S. Using proxies for OOV keywords in the keyword search task. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2013), Olomouc, Czech Republic, 8–12 December 2013; pp. 416–421.
79. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2016, Shanghai, China, 20–25 March 2016; pp. 4960–4964.
80. Ali, A.; Clements, M.A. Spoken Web Search using an Ergodic Hidden Markov Model of Speech. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013; pp. 861–862.
81. Caranica, A.; Buzo, A.; Cucu, H.; Burileanu, C. Speed@MediaEval 2015: Multilingual Phone Recognition Approach to Query by Example STD. In Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14–15 September 2015; pp. 781–783.
82. Kesiraju, S.; Mantena, G.; Prahallad, K. IIIT-H System for MediaEval 2014 QUESST. In Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, 16–17 October 2014; pp. 761–762.
83. Ma, M.; Rosenberg, A. CUNY Systems for the Query-by-Example Search on Speech Task at MediaEval 2015. In Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14–15 September 2015; pp. 831–833.
84. Takahashi, J.; Hashimoto, T.; Konno, R.; Sugawara, S.; Ouchi, K.; Oshima, S.; Akyu, T.; Itoh, Y. An IWAPU STD System for OOV Query Terms and Spoken Queries. In Proceedings of the 11th NTCIR Workshop, Tokyo, Japan, 9–12 December 2014; pp. 384–389.
85. Makino, M.; Kai, A. Combining Subword and State-level Dissimilarity Measures for Improved Spoken Term Detection in NTCIR-11 SpokenQuery&Doc Task. In Proceedings of the 11th NTCIR Workshop, Tokyo, Japan, 9–12 December 2014; pp. 413–418.
86. Sakamoto, N.; Yamamoto, K.; Nakagawa, S. Combination of syllable based N-gram search and word search for spoken term detection through spoken queries and IV/OOV classification. In Proceedings of the 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Scottsdale, AZ, USA, 13–17 December 2015; pp. 200–206.
87. Hou, J.; Pham, V.T.; Leung, C.C.; Wang, L.; Xu, H.; Lv, H.; Xie, L.; Fu, Z.; Ni, C.; Xiao, X.; et al. The NNI Query-by-Example System for MediaEval 2015. In Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14–15 September 2015; pp. 141–143.
88. Vavrek, J.; Vizslay, P.; Lojka, M.; Pleva, M.; Juhar, J.; Rusko, M. TUKE at MediaEval 2015 QUESST. In Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14–15 September 2015; pp. 451–453.
89. Wang, H.; Lee, T.; Leung, C.C.; Ma, B.; Li, H. Acoustic Segment Modeling with Spectral Clustering Methods. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 264–277. [[CrossRef](#)]
90. Chung, C.T.; Lee, L.S. Unsupervised discovery of structured acoustic tokens with applications to spoken term detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 394–405. [[CrossRef](#)]
91. Chung, C.T.; Tsai, C.Y.; Liu, C.H.; Lee, L.S. Unsupervised iterative Deep Learning of speech features and acoustic tokens with applications to spoken term detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1914–1928. [[CrossRef](#)]
92. Ram, D.; Miculicich, L.; Boulard, H. Multilingual Bottleneck Features for Query by Example Spoken Term Detection. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, 14–18 December 2019; pp. 621–628.
93. Ram, D.; Miculicich, L.; Boulard, H. Neural Network Based End-to-End Query by Example Spoken Term Detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1416–1427. [[CrossRef](#)]
94. Hazen, T.J.; Shen, W.; White, C.M. Query-by-Example spoken term detection using phonetic posteriorgram templates. In Proceedings of the Eleventh Biannual IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Merano, Italy, 13–17 December 2009; pp. 421–426.
95. Tulsiani, H.; Rao, P. The IIT-B Query-by-Example System for MediaEval 2015. In Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14–15 September 2015; pp. 341–343.
96. Proenca, J.; Veiga, A.; Perdigão, F. The SPL-IT Query by Example Search on Speech system for MediaEval 2014. In Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, 16–17 October 2014; pp. 741–742.
97. Proenca, J.; Veiga, A.; Perdigão, F. Query by example search with segmented dynamic time warping for non-exact spoken queries. In Proceedings of the 23rd European Signal Processing Conference, Nice, France, 31 August–4 September 2015; pp. 1691–1695.
98. Proenca, J.; Castela, L.; Perdigão, F. The SPL-IT-UC Query by Example Search on Speech system for MediaEval 2015. In Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14–15 September 2015; pp. 471–473.
99. Proenca, J.; Perdigão, F. Segmented Dynamic Time Warping for Spoken Query-by-Example Search. In Proceedings of the Interspeech, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 750–754.
100. Lopez-Otero, P.; Docio-Fernandez, L.; Garcia-Mateo, C. GTM-UVigo Systems for the Query-by-Example Search on Speech Task at MediaEval 2015. In Proceedings of the 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Scottsdale, AZ, USA, 13–17 December 2015; pp. 521–523.
101. Lopez-Otero, P.; Docio-Fernandez, L.; Garcia-Mateo, C. Phonetic Unit Selection for Cross-Lingual Query-by-Example Spoken Term Detection. In Proceedings of the 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Scottsdale, AZ, USA, 13–17 December 2015; pp. 223–229.

102. Saxena, A.; Yegnanarayana, B. Distinctive Feature Based Representation of Speech for Query-by-Example Spoken Term Detection. In Proceedings of the Interspeech, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp. 3680–3684.
103. Lopez-Otero, P.; Docio-Fernandez, L.; Garcia-Mateo, C. Compensating Gender Variability in Query-by-Example Search on Speech Using Voice Conversion. In Proceedings of the Interspeech, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 2909–2913.
104. Asaei, A.; Ram, D.; Boulard, H. Phonological Posterior Hashing for Query by Example Spoken Term Detection. In Proceedings of the Interspeech, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018; pp. 2067–2071.
105. Mantena, G.; Achanta, S.; Prahallad, K. Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 946–955. [[CrossRef](#)]
106. Wang, H.; Lee, T. The CUHK Spoken Web Search System for MediaEval 2013. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013; pp. 681–682.
107. Wang, H.; Lee, T.; Leung, C.C.; Ma, B.; Li, H. Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection. In Proceedings of the 38th International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013, Vancouver, BC, Canada, 26–31 May 2013; pp. 8545–8549.
108. Torbati, A.H.H.N.; Picone, J. A nonparametric bayesian approach for spoken term detection by example query. In Proceedings of the Interspeech, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 928–932.
109. Popli, A.; Kumar, A. Query-by-Example Spoken Term Detection Using Low Dimensional Posteriorgrams Motivated by Articulatory Classes. In Proceedings of the 17th IEEE International Workshop on Multimedia Signal Processing, MMSP 2015, Xiamen, China, 19–21 October 2015; pp. 1–6.
110. Ram, D.; Asaei, A.; Boulard, H. Sparse Subspace Modeling for Query by Example Spoken Term Detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1130–1143. [[CrossRef](#)]
111. Skacel, M.; Szöke, I. BUT QUESST 2015 System Description. In Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14–15 September 2015; pp. 721–723.
112. Chen, H.; Leung, C.C.; Xie, L.; Ma, B.; Li, H. Unsupervised Bottleneck Features for Low-Resource Query-by-Example Spoken Term Detection. In Proceedings of the Interspeech, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 923–927.
113. Yuan, Y.; Leung, C.C.; Xie, L.; Chen, H.; Ma, B.; Li, H. Pairwise learning using multi-lingual bottleneck features for low-resource query-by-example spoken term detection. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 May 2017; pp. 5645–5649.
114. van Hout, J.; Mitra, V.; Franco, H.; Bartels, C.; Vergyri, D. Tackling unseen acoustic conditions in query-by-example search using time and frequency convolution for multilingual deep bottleneck features. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 48–54.
115. Yilmaz, E.; van Hout, J.; Franco, H. Noise-robust exemplar matching for rescoring query-by-example search. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 1–7.
116. Bouallegue, M.; Senay, G.; Morchid, M.; Matrouf, D.; Linares, G.; Dufour, R. LIA@MediaEval 2013 Spoken Web Search Task: An I-Vector based Approach. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013; pp. 771–772.
117. Rodriguez-Fuentes, L.J.; Varona, A.; Penagarikano, M.; Bordel, G.; Diez, M. GTTS Systems for the SWS Task at MediaEval 2013. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013; pp. 831–832.
118. Yang, P.; Leung, C.C.; Xie, L.; Ma, B.; Li, H. Intrinsic Spectral Analysis based on temporal context features for Query-by-Example Spoken Term Detection. In Proceedings of the Interspeech, 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; pp. 1722–1726.
119. George, B.; Saxena, A.; Mantena, G.; Prahallad, K.; Yegnanarayana, B. Unsupervised Query-by-Example Spoken Term Detection using Bag of Acoustic Words and Non-segmental Dynamic Time Warping. In Proceedings of the Interspeech, 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; pp. 1742–1746.
120. Zhan, J.; He, Q.; Su, J.; Li, Y. A Stage Match for Query-by-Example Spoken Term Detection Based On Structure Information of Query. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, 6–11 June 2021; pp. 6833–6837.
121. Abad, A.; Astudillo, R.F.; Trancoso, I. The L2F Spoken Web Search system for Mediaeval 2013. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013; pp. 851–852.
122. Szöke, I.; Skácel, M.; Burget, L. BUT QUESST 2014 System Description. In Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, 16–17 October 2014; pp. 621–622.
123. Szöke, I.; Burget, L.; Grézl, F.; Černocký, J.H.; Ondel, L. Calibration and fusion of query-by-example systems—BUT SWS 2013. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 7849–7853.

124. Abad, A.; Rodríguez-Fuentes, L.J.; Penagarikano, M.; Varona, A.; Bordel, G. On the calibration and fusion of heterogeneous spoken term detection systems. In Proceedings of the Interspeech, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013; pp. 20–24.
125. Yang, P.; Xu, H.; Xiao, X.; Xie, L.; Leung, C.C.; Chen, H.; Yu, J.; Lv, H.; Wang, L.; Leow, S.J.; et al. The NNI Query-by-Example System for MediaEval 2014. In Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, 16–17 October 2014; pp. 691–692.
126. Leung, C.C.; Wang, L.; Xu, H.; Hou, J.; Pham, V.T.; Lv, H.; Xie, L.; Xiao, X.; Ni, C.; Ma, B.; et al. Toward High-Performance Language-Independent Query-by-Example Spoken Term Detection for MediaEval 2015: Post-Evaluation Analysis. In Proceedings of the Interspeech, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 3703–3707.
127. Xu, H.; Hou, J.; Xiao, X.; Pham, V.T.; Leung, C.C.; Wang, L.; Do, V.H.; Lv, H.; Xie, L.; Ma, B.; et al. Approximate search of audio queries by using DTW with phone time boundary and data augmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2016, Shanghai, China, 20–25 March 2016; pp. 6030–6034.
128. Oishi, S.; Matsuba, T.; Makino, M.; Kai, A. Combining State-level and DNN-based Acoustic Matches for Efficient Spoken Term Detection in NTCIR-12 SpokenQuery&Doc-2 Task. In Proceedings of the 12th NTCIR Workshop, Tokyo, Japan, 7–10 June 2016; pp. 205–210.
129. Obara, M.; Kojima, K.; Tanaka, K.; wook Lee, S.; Itoh, Y. Rescoring by Combination of Posteriorgram Score and Subword-Matching Score for Use in Query-by-Example. In Proceedings of the Interspeech, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 1918–1922.
130. Ram, D.; Miculicich, L.; Boursard, H. CNN Based Query by Example Spoken Term Detection. In Proceedings of the Interspeech, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018; pp. 92–96.
131. Shankar, R.; Vikram, C.M.; Prasanna, S.M. Spoken Keyword Detection Using Joint DTW-CNN. In Proceedings of the Interspeech, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018; pp. 117–121.
132. Settle, S.; Levin, K.; Kamper, H.; Livescu, K. Query-by-Example Search with Discriminative Neural Acoustic Word Embeddings. In Proceedings of the Interspeech, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 2874–2878.
133. Yuan, Y.; Leung, C.C.; Xie, L.; Chen, H.; Ma, B.; Li, H. Learning Acoustic Word Embeddings with Temporal Context for Query-by-Example Speech Search. In Proceedings of the Interspeech, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018; pp. 97–101.
134. Zhu, Z.; Wu, Z.; Li, R.; Meng, H.; Cai, L. Siamese Recurrent Auto-Encoder Representation for Query-by-Example Spoken Term Detection. In Proceedings of the Interspeech, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018; pp. 102–106.
135. Ao, C.W.; Lee, H.Y. Query-by-example spoken term detection using attention-based multi-hop networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, 15–20 April 2018; pp. 6264–6268.
136. Santos, C.D.; Tan, M.; Xiang, B.; Zhou, B. Attentive pooling networks. *arXiv* **2016**, arXiv:1602.03609.
137. Zhang, K.; Wu, Z.; Jia, J.; Meng, H.; Song, B. Query-by-example spoken term detection using attentive pooling networks. In Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 1267–1272.
138. NIST. The Spoken Term Detection (STD) 2006 Evaluation Plan. 2006. Available online: <https://catalog.ldc.upenn.edu/docs/LDC2011S02/std06-evalplan-v10.pdf> (accessed on 10 september 2021).
139. ITU. *ITU-T Recommendation P.563: Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications*; ITU: Geneva, Switzerland, 2008.
140. Lleida, E.; Ortega, A.; Miguel, A.; Bazán, V.; Pérez, C.; Zotano, M.; de Prada, A. *RTVE2018 Database Description*; Vivolab and Corporación Radiotelevisión Española: Zaragoza, Spain, 2018.
141. Martin, A.; Doddington, G.; Kamm, T.; Ordowski, M.; Przybocki, M. The DET Curve In Assessment Of Detection Task Performance. In Proceedings of the 5th European Conference on Speech Communication and Technology, Rhodes, Greece, 22–25 September 1997; pp. 1895–1898.
142. NIST. *Evaluation Toolkit (STDEval) Software*; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 1996.
143. Akiba, T.; Nishizaki, H.; Nanjo, H.; Jones, G.J.F. Overview of the NTCIR-11 SpokenQuery&Doc Task. In Proceedings of the 11th NTCIR Workshop, Tokyo, Japan, 9–12 December 2014; pp. 1–15.
144. Akiba, T.; Nishizaki, H.; Nanjo, H.; Jones, G.J.F. Overview of the NTCIR-12 SpokenQuery&Doc-2 Task. In Proceedings of the 12th NTCIR Workshop, Tokyo, Japan, 7–10 June 2016; pp. 1–13.
145. Fiscus, J.; Ajot, J.; Doddington, G. *English STD 2006 Results*; Technical Report; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2006.
146. Harper, M.P. Data Resources to Support the Babel Program Intelligence Advanced Research Projects Activity (IARPA). 2011. Available online: <https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/harper.pdf> (accessed on 10 september 2021).

147. Hartmann, W.; Karakos, D.; Hsiao, R.; Zhang, L.; Alumae, T.; Tsakalidis, S.; Schwartz, R. Analysis of keyword spotting performance across IARPA Babel languages. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 May 2017; pp. 5765–5769.
148. Sailor, H.B.; Patil, A.T.; Patil, H.A. Advances in Low Resource ASR: A Deep Learning Perspective. In Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU'18), Gurugram, India, 29–31 August 2018; pp. 15–19.
149. Ragni, A.; Li, Q.; Gales, M.J.F.; Wang, Y. Confidence Estimation and Deletion Prediction Using Bidirectional Recurrent Neural Networks. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 204–211.
150. Karafiat, M.; Baskar, M.K.; Vesely, K.; Grezl, F.; Burget, L.; Cernocky, J. Analysis of Multilingual BLSTM Acoustic Model on Low and High Resource Languages. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, 15–20 April 2018; pp. 5789–5793.
151. Yusuf, B.; Saraclar, M. An Empirical Evaluation of DTW Subsampling Methods for Keyword Search. In Proceedings of the Interspeech, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019; pp. 2673–2677.
152. Yusuf, B.; Gundogdu, B.; Saraclar, M. Low Resource Keyword Search With Synthesized Crosslingual Exemplars. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1126–1135. [[CrossRef](#)]
153. Piunova, A.; Beck, E.; Schluter, R.; Ney, H. Rescoring Keyword Search Confidence Estimates with Graph-based Re-ranking Using Acoustic Word Embeddings. In Proceedings of the Interspeech, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019; pp. 4205–4209.
154. Yi, J.; Tao, J.; Bai, Y. Language-invariant Bottleneck Features from Adversarial End-to-end Acoustic Models for Low Resource Speech Recognition. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, UK, 12–17 May 2019; pp. 6071–6075.
155. Yi, J.; Tao, J.; Wen, Z.; Bai, Y. Language-Adversarial Transfer Learning for Low-Resource Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 621–630. [[CrossRef](#)]
156. Gundogdu, B.; Yusuf, B.; Saraclar, M. Generative RNNs for OOV Keyword Search. *IEEE Signal Process. Lett.* **2019**, *26*, 124–128. [[CrossRef](#)]
157. Gök, A.; Gündoğdu, B.; Saraçlar, M. Accelerating Dynamic Time Warping-Based Keyword Search Using Recurrent Neural Networks. In Proceedings of the 28th IEEE Conference on Signal Processing and Communications Applications, Gaziantep, Turkey, 5–7 October 2020.
158. Fantaye, T.G.; Yu, J.; Hailu, T.T. Advanced Convolutional Neural Network-Based Hybrid Acoustic Models for Low-Resource Speech Recognition. *Computers* **2020**, *9*, 36. [[CrossRef](#)]
159. Thomas, S.; Audhkhasi, K.; Kingsbury, B. Transliteration Based Data Augmentation for Training Multilingual ASR Acoustic Models in Low Resource Settings. In Proceedings of the Interspeech, 21st Annual Conference of the International Speech Communication Association, Shanghai, China, 25–29 October 2020; pp. 4736–4740.
160. NIST. *OpenKWS13 Keyword Search Evaluation Plan*; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2013.
161. NIST. *Draft KWS14 Keyword Search Evaluation Plan*; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2013.
162. NIST. *KWS15 Keyword Search Evaluation Plan*; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2015.
163. NIST. *Draft KWS16 Keyword Search Evaluation Plan*; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2016.
164. NIST. *Open Speech Analytic Technologies Pilot Evaluation*; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2018.
165. NIST. *2017 Pilot Open Speech Analytic Technologies Evaluation (2017 NIST Pilot OpenSAT) Post Evaluation Summary*; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2019.
166. NIST. *NIST Open Speech Analytic Technologies 2019 Evaluation Plan (OpenSAT19)*; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2019.
167. NIST. *NIST Open Speech Analytic Technologies 2020 Evaluation Plan (OpenSAT20)*; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2020.
168. Rajput, N.; Metze, F. Spoken Web Search. In Proceedings of the MediaEval 2011 Workshop, Pisa, Italy, 1–2 September 2011; pp. 1–2.
169. Metze, F.; Barnard, E.; Davel, M.; van Heerden, C.; Anguera, X.; Gravier, G.; Rajput, N. The Spoken Web Search Task. In Proceedings of the MediaEval 2012 Workshop, Pisa, Italy, 4–5 October 2012; pp. 41–42.
170. Anguera, X.; Metze, F.; Buzo, A.; Szoke, I.; Rodriguez-Fuentes, L.J. The Spoken Web Search Task. In Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, 18–19 October 2013; pp. 1–2.
171. Povey, D. Kaldi-ASR. 2014. Available online: <https://github.com/kaldi-asr/kaldi/tree/master/egs/ws/s5> (accessed on 10 september 2021).

172. Povey, D.; Burget, L.; Agarwal, M.; Akyazi, P.; Kai, F.; Ghoshal, A.; Glembek, O.; Goel, N.; Karafiat, M.; Rastrow, A.; et al. The subspace Gaussian mixture model: A structured model for speech recognition. *Comput. Speech Lang.* **2011**, *25*, 404–439. [[CrossRef](#)]
173. Stolcke, A. SRILM—An Extensible Language Modeling Toolkit. In Proceedings of the Interspeech, 7th International Conference on Spoken Language Processing, ICSLP2002, Denver, CO, USA, 16–20 September 2002; pp. 901–904.
174. Silva, J.C. Multilingual Grapheme to Phoneme. 2015. Available online: <https://github.com/jcsilva/multilingual-g2p> (accessed on 10 september 2021).
175. Can, D.; Saraclar, M. Lattice indexing for spoken term detection. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2338–2347. [[CrossRef](#)]
176. Thiemann, J.; Ito, N.; Vincent, E. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings. In Proceedings of the 21st International Congress on Acoustics, Montréal, QC, Canada, 2–7 June 2013; p. 035081.
177. Hirsch, H.G. Fant-Filtering and Noise Adding Tool. Niederrhein University of Applied Sciences. 2005. Available online: <http://dnt.kr.hs-niederrhein.de/indexbd2f.html> (accessed on 10 september 2021).
178. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the 40th International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, Brisbane, Australia, 19–24 April 2015; pp. 5206–5210.